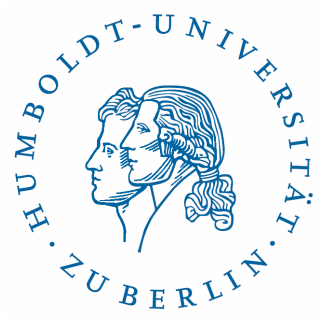

Extreme value analysis of speeding data

Bachelor's Thesis submitted

to

Prof. Dr. W.K. Härdle

Humboldt-Universität zu Berlin
School of Business and Economics
Institute for Statistics and Econometrics
Ladislaus von Bortkiewicz Chair of Statistics

by

Alexander Buchholz
(532 847)

in partial fulfillment of the requirements
for the degree of
Bachelor of Science in Economics
Paris, 4th of October, 2013

Acknowledgement

First of all, I would like to thank Prof. Dr. Wolfgang K. Härdle for supervising this Bachelor's thesis, for his quick responses to my questions and for his helpful remarks. Also, I would like to thank Dr. Julia Schaumburg, without whom this thesis would not have been possible. Especially, I would like to thank her for her mentoring, for encouraging me in writing this thesis and for her helpful advice.

Furthermore, I would like to thank Prof. Dr. Melanie Schienle and Prof. Dr. Nikolaus Hautsch. Due to their recommendations, I was able to realize my studies in Paris.

I would like to thank my friends Audrey Thenot, Emilien Macault, Guillaume Meyer, Leo Pape, Friederike Berlinghoff, Georg Bieker and Andre Jüling for their helpful remarks and comments on this thesis. I would like to thank my family, without whom I would never have reached the point at which I am now. Finally, I would like to thank Birgit Schaffer for being there for me anytime I need her.

Abstract

Is extreme value theory a suitable approach for modeling the behavior of speeding data? In the following thesis I will reply to this question by introducing the basic concepts of extreme value theory. For this purpose, I am going to analyze a set of speeding values recorded by the Berlin police from 2009 to 2011. First, I will approach this question by using basic statistic indicators. Afterwards, I am going to enlarge the understanding of the underlying distribution by using a quantile plot approach. Finally, I will calculate the extreme value index by using Hill's estimator and moment estimators as well as the endpoint. I am going to show that the resulting endpoint depends on the chosen estimator for the extreme value index. For several speed limits, the realized maximum will be close to the estimated endpoint, whereas for other speed classes an even more excessive speeding behavior should be expected.

Keywords: speeding data, extreme value theory, Fréchet, Gumbel and Weibull distribution, quantile regression, extreme value index, endpoint estimation

Contents

List of Figures

List of Tables

1 Introduction

In our daily life, we are often interested in the average behavior of characteristic numbers. For example, what is the average waiting time for a bus? What is the average return of a financial investment? And what is the average human life expectation? Extreme values in our day to day life are often regarded as exceptional events that are rare and less relevant. However, the characterization of extremal events is often far more crucial than the average behavior of any indicator. What maximum daily loss might occur on the stock market? What is the maximum claim size an insurance company should be prepared to handle to avoid bankruptcy? And how high a dike should be built to be an effective protection against floods, not only given measurements of the last 100 years, but for every flood one should expect, given todays climate conditions? When dealing with such extremal events, ordinary statistics based on the central limit theorem fail. At this point, extreme value theory comes into play and provides the necessary concepts to deal with extremely rare events. As these extremal events might cause high damage in both an economic and human perspective, extreme value theory experienced growing academic and practical interest. Various applications such as in meteorology (e.g. Palutikof et al. (1999)), risk management (e.g. Embrechts et al. (1999)), finance and insurance (e.g. Embrechts et al. (1997)) and even world records in sports (see for example Einmahl and Magnus (2008)) have emerged.

The question I want to examine in the following thesis is, how one can apply these methods to traffic data, especially the observation of speeding data, recorded by the Berlin police, Germany, from 2009 to 2011. I will principally follow the method of Einmahl and Magnus (2008), who estimated the endpoint of world records in athletics and the quality of current records using extreme value theory. For this purpose, I will motivate the use of extreme value theory by showing that an approach by Gaussian modeling is not satisfying. Then an independent modeling by the three extremal type distributions will support the introduction of extreme value theory. Furthermore, the estimation of the extreme value index will be necessary to understand the tail behavior of the speeding data distributions. Finally, the main question I want to investigate is, what the potential endpoint of the speeding distributions might be. This topic is of a high importance because ex ante it is not clear whether such a finite endpoint exists or not. In this way, I want to find out whether there is a natural boundary to human speeding behavior, or if today's speeding behavior is only limited by the given vehicle construction and local road conditions and even more excessive speeding behavior should be expected. I will extend the analysis grouped by speed limits to a grouping

by months and by half-years. This regrouping will allow me to potentially identify seasonal behavior and a changing of excess speeding over time.

I am going to show that the speeding data can be modeled best by a Weibull distribution and that the Gumbel distribution represents a fair fit as well. The modeling by a Fréchet distribution is less adequate. This will imply the existence of a finite endpoint for all speed limits. However, this modeling will depend on the chosen estimator for the extreme value index. The value of the endpoint depends on this choice, as well. Besides, I will show that there might be seasonal effects in the speeding behavior. Basically, in summer and winter the speeding is less excessive than in fall and spring. On the other hand, a general increase or decrease of the speeding behavior over time will not be confirmed by the methods used.

This bachelor-thesis is organized in the following way: first, I will introduce the dataset for my analysis and give basic statistical results. This preliminary overview of the data already motivates the use of extreme value theory. By using box plots, histograms and kernel density estimations, I will show that an adequate fit of the data can be modeled by the three extreme value distributions: the Fréchet, Weibull and Gumbel distribution. The use of these distributions will be justified by a quantile approach using quantile regression and quantile-quantile-plots in section 3. These graphical tools will deepen the understanding of the data and furthermore support an extreme value theory approach. In section 4, I am going to detail the theoretical part of extreme value theory that will be used for my question. For this purpose, I will explain the estimators for the extreme value index γ and the method for the endpoint estimation, x^*. In section 5 I will give the results of the extreme value index and endpoint estimations and detail the procedure I used. Finally, I will conclude with section 6, giving further ideas on how to improve my procedure and suggest other questions that one could analyze by using the given dataset.

2 Data

2.1 Description of the dataset

The dataset I will analyze in the following, is from the Berlin police, Germany, and contains the speeding excess values of mobile speed measurement stations for the entire years 2009, 2010, and 2011. It contains for every speed measurement the time and date of the measurement period, the number of cars captured, the allowed speed-limit, the number of cars exceeding the speed limit, the highest speed measured, the number of fees, penalties and suspended driving licenses. In the following, I am basically interested in the behavior of these highest measured speeds during one measurement period. So even if there were several cars exceeding the speed limit, I will only consider the absolute maximum for a measurement period. Altogether there are 40453 measuring points that are divided into 11 speed limit classes from 7 km/h to 130 km/h. For the highest speed limit of 130 km/h, there are only two exploitable values, so I exclude these from the analysis. For 100 km/h, there are also only 8 observations, so I exclude this speed limit from this analysis as well. This reduces the dataset to 40443 observations. There is no information about the cars that were recorded. So it might be possible that an individual appears several times in the dataset, having exceeded the speed limit, for instance, twice a day or several times a year. Let $i \in \{1, 2, ..., 40443\}$ be an index running through the number of observations. Let $j \in \{7, 10, 20, 30, 40, 50, 60, 70, 80\}$ be an index that runs through the different speed limits. Consequently, i, j denotes the subset of speeding values for a given speed limit. Since the maximum of the measured speed in the initial dataset is always an integer, I introduce a smoothing-method in order to simulate the uncertainty that arises while measuring the speed. For this purpose, I add a random uniform number between -0.5 and 0.5 to the values. This smoothing-method was necessary, otherwise I would have to deal with multiple values in clusters that could cause problems for the estimations I want to calculate in the following (see (Einmahl and Magnus, 2008, p. 1383)). In order to normalize the values, I subtract the allowed speed limit from every measured speed maximum. This allows me to analyze the excess speed with respect to the underlying speed limit and also for the pooled observations. Let $Z_{i,j}$ be the random maximum speed observed. Let $u_{i,j} \sim \mathbb{U}[-0.5, 0.5]$ for $i \in \{1, 2, ..., 40443\}$ and $j \in \{7, 10, 20, 30, 40, 50, 60, 70, 80\}$ be identically and independently distributed random uniform numbers and let SL_j be the speed limit for j. I then define the smoothed and normed speeding values by

$$X_{i,j} = Z_{i,j} - SL_j + u_{i,j}. \tag{2.1}$$

Also I define the following order statistic: $X_{1_j,n_j}, X_{2_j,n_j}, ..., X_{n_j,n_j}$, where X_{n_j,n_j} denotes the absolute maximum for the speed class j. In the following, I will always refer to this smoothed and normalized values as the speed excess values in the analysis.

2.2 Basic statistical analysis

Even when the type of the data immediately suggests the use of methods for extreme value analysis, I want to justify this approach by taking a closer look on the data. For this purpose, I will first consider basic statistic indicators like location and scale parameters. Moreover, I would like to use graphical tools to better explain the homogeneity or heterogeneity of the different speed classes.

Speed limit in km/h	7	10	20	30	40	50	60	70	80	overall
Number of observations	560	266	125	17914	77	20530	641	96	234	40443
Minimum	4.24	7.48	7.92	2.49	7.98	1.53	1.20	6.33	17.39	1.20
1st quartile	17.24	16.30	21.20	16.44	17.86	16.74	23.31	18.77	34.08	16.69
Median	21.80	20.83	25.21	20.81	26.49	21.75	29.66	24.66	41.25	21.45
Mean	21.70	20.85	25.33	21.86	29.43	23.45	31.27	25.88	42.88	22.96
3rd quartile	25.90	24.40	28.82	26.06	37.21	28.17	36.81	31.07	49.26	27.42
Maximum	42.96	42.05	46.30	111.00	77.43	98.05	100.10	68.69	98.66	111.00
Standard Deviation	6.18	6.35	6.81	7.85	14.10	9.61	12.49	10.38	13.85	9.15
Skewness	0.14	0.43	0.41	1.19	0.84	1.32	1.36	1.14	1.02	1.44

Table 2.1: Descriptive statistics for 40443 observed speed maxima (smoothed and normalized) for the period from January 01, 2011 to December 31, 2011 recorded by the Berlin police. Numbers are rounded to two decimal points.

Table 2.1 contains a first overview of the observations. I get the impression that a higher speed limit tends to have higher measured maxima. Furthermore, the volatility of the data also tends to augment. On the other hand, these effects seem to be ambiguous when taking a look at the speed classes for 30 and 50 km/h, containing the most observations. Along with higher absolute maxima for the different speed classes, I get larger skew. Thus, extremal events become more likely. Taking a closer look at the box plot in figure 2.1, underlines this impression. In order to understand the differences between the speed limits, I will analyze both the overall speeding (as pooled data) and the speeding grouped by their speed limit class.

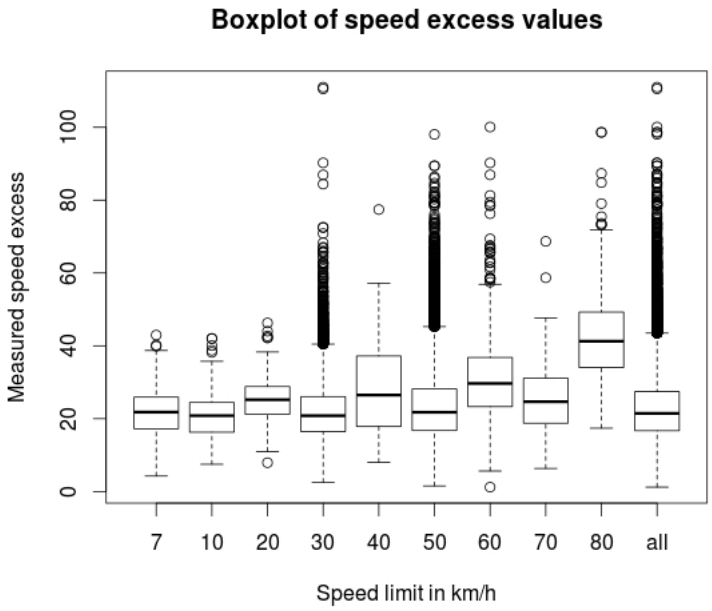

Figure 2.1: Box plot of smoothed and normalized speeding values in km/h for the different speed limits.

2.3 Histograms and density estimations

After the presentation of these basic statistic indicators, I will introduce the functions that represent the essential distribution functions in extreme value theory: the Gumbel distribution, the Weibull distribution and the Fréchet distribution. But at this point one may still ask what convinces me to consider my question as a problem that needs extreme value theory. I will reply to that question by taking a look at possible, but not satisfying alternatives.
The central limit theorem represents a powerful tool for the explanation of the limit behavior of the sum of random variables if the number of observations is large enough. Furthermore, the Gaussian distribution is widely used to approximate the behavior of unknown random variables. Hence, it seems natural to model the behavior of the speed maxima by a Gaussian distribution, using moment estimators for the scale parameter σ and the location parameter μ . On the other hand, I know that when taking only maxima of a sample (in this case the observed speed values), the underlying distribution tends either to a Fréchet, a Gumbel or a Weibull distribution. This is the central result of the extremal types theorem that I will study in section 4. In this way, it would be a natural approach to estimate the parameters of the extremal types distributions and compare them to Gaussian distributions and kernel density estimations. However, this approach proved to be inefficient, since the estimation of

the distribution parameters came along with numerical problems. Either the use of moment estimators or of maximum likelihood estimators failed using different `R` packages. Therefore, I use regression based estimations exploiting the quantile plots of section 3, that give additional information on the underlying distributions. The parameters I used for the density plots are taken from these estimations (see table 3.1 for the shape and location parameters of the Weibull and Fréchet distributions). Nevertheless, I will still explain the basic idea of the maximum likelihood estimations and use moment estimators for the Gumbel and Gaussian distribution. In section 3 and 4, I will derive theoretically the statistical tools for extreme value theory that I already use at this point.

2.3.1 The normal distribution

The Gaussian distribution, given by its density function

$$f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \tag{2.2}$$

with the domain $x \in \mathbb{R}$ and $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*$, is the first distribution I want to fit to the speeding data. See (Bertsekas and Tsitsiklis, 2002, p. 152-158) for further information on the Gaussian distribution. By using the moment estimators from table 2.1, I get the following Gaussian model for every speed class: μ_j =mean, σ_j =standard deviation and thus, $X_j \sim \mathcal{N}(\mu_j, \sigma_j)$. μ_j is estimated by $\frac{\sum_{i=1}^n x_{i,j}}{n}$ and σ_j is estimated by $\sqrt{\frac{1}{n-1}\sum_{i=1}^n (x_{i,j} - \overline{x_j})^2}$. However, the idea of supposing a Gaussian model is clearly only a very bad approximation, since I have a clear left endpoint given by 0, since only the speed exceeding the speed limit is recorded.

2.3.2 The Fréchet distribution

For the Fréchet distribution, the first idea would be to use a maximum-likelihood approach to estimate the parameters of the density function. But this method failed due to numerical calculation problems and showed to be biased. See Mubarak (2011) for details on the maximum-likelihood approach, for example. Another approach, that I am not going to study, was proposed by Gumbel (1965). See there for further information on the Fréchet distribution.

The basic procedure, one would use for a maximum likelihood approach, would be the following: if a random number X_i follows a Fréchet distribution with scale parameter s and shape parameter α (the location m is set to 0), I then have the following density function f

with the domain $x \geq 0$ and $(s, \alpha) \in \mathbb{R}_+^{*2}$ and hence, get the likelihood-function V for a set of independent identically distributed random variables:

$$f_{\alpha,s}(x_i) = \frac{\alpha}{s}\left(\frac{x_i}{s}\right)^{-1-\alpha} \exp\left\{-\left(\frac{x_i}{s}\right)^{-\alpha}\right\}, \tag{2.3}$$

$$V_{\alpha,s}(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} \frac{\alpha}{s}\left(\frac{x_i}{s}\right)^{-1-\alpha} \exp\left\{-\left(\frac{x_i}{s}\right)^{-\alpha}\right\}, \tag{2.4}$$

and by taking the negative logarithm I get:

$$-\ln(V_{\alpha,s}(x_1, x_2, ..., x_n)) = \sum_{i=1}^{n}\left[(1+\alpha)\ln\left(\frac{x_i}{s}\right) - \ln\left(\frac{\alpha}{s}\right) + \left(\frac{x_i}{s}\right)^{-\alpha}\right]. \tag{2.5}$$

I minimize the negative log-likelihood-function in order to obtain the maximum-likelihood (ml) estimations for α and s. This leads to

$$-\frac{\partial \ln(V_{\alpha,s}(x_1, x_2, ..., x_n))}{\partial \alpha} = \sum_{i=1}^{n}\left[\ln\left(\frac{x_i}{s}\right) - \frac{1}{\alpha} - \ln\left(\frac{x_i}{s}\right)\left(\frac{x_i}{s}\right)^{-\alpha}\right] \stackrel{!}{=} 0$$

$$-\frac{\partial \ln(V_{\alpha,s}(x_1, x_2, ..., x_n))}{\partial s} = \sum_{i=1}^{n}\left[(1+\alpha)\left(\frac{-x_i}{s^2}\right) + \frac{1}{s} + \alpha x_i^{-\alpha} s^{\alpha-1}\right] \stackrel{!}{=} 0$$

Because the calculation of these parameters is a non-trivial numerical problem, I tried to use the package `VGAM` and the function `Frechet2` available for the statistical software `R` to calculate the ml-estimators for the Fréchet-distribution (see Yee (2013)). However, this approach did not yield any results. Instead, I used the results from the quantile-regression procedure that can be found in table 3.1. Since the Fréchet distribution has a slowly decaying tail, no finite endpoint exists and even faster excess speeding would be attributed a positive probability. If this model represents a good fit to the data, it will be impossible to estimate a global speed maximum.

2.3.3 The Gumbel distribution

For the Gumbel-distribution, with the domain $x \in \mathbb{R}$ and $(\mu, s) \in \mathbb{R} \times \mathbb{R}_+^*$, which has the density-function f, with scale parameter s and location parameter μ,

$$f_{\mu,s}(x) = \frac{1}{s}\exp\left[-\frac{x-\mu}{s} - \exp\left(-\frac{x-\mu}{s}\right)\right], \tag{2.6}$$

I use simple moment estimators to find the unknown parameters. The first is the empirical expectation given by

$$\overline{X} = \sum_{i=1}^{n}\frac{X_i}{n} \to^P \mathbb{E}[X] = \mu + s\psi,$$

where $\psi \approx 0.5772$ denotes the Euler-Mascheroni constant. The second is the empirical variance given by

$$\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \to^P \mathbb{V}ar\left[X\right] = \frac{\pi^2 s^2}{6}.$$

The convergence in probability is derived by the weak law of large numbers (see (Bertsekas and Tsitsiklis, 2002, p. 384). This yields a simple estimation for the two unknown parameters s and μ given by

$$\overline{X} - \frac{\widehat{\sigma}\sqrt{6}}{\pi}\psi \approx \mu, \qquad \frac{\widehat{\sigma}\sqrt{6}}{\pi} \approx s. \tag{2.7}$$

This approach is taken from (Landwehr et al., 1979, p. 1056-1057). The estimations for those parameters can be found in table 2.2. If the data can be fitted well by a Gumbel distribution, this would be an indicator for a non-finite endpoint, thus faster excess speeding would have a positive probability. But due to the exponential decay of the the tail, less extremal events should be observed compared to a Fréchet distribution.

Speed limit in km/h	7	10	20	30	40	50	60	70	80	overall
μ	18.95	18.00	22.27	18.33	23.08	19.12	25.65	21.21	36.65	18.85
s	4.77	4.95	5.31	6.12	10.99	7.49	9.74	8.10	10.80	7.13

Table 2.2: Location and scale parameter estimations for the Gumbel distribution. Numbers are rounded to two decimal points. Estimation errors are omitted.

2.3.4 The Weibull distribution

For the Weibull distribution, I tried to use once more a maximum-likelihood approach, as proposed by Smith and Naylor (1987). The Weibull distribution, with the domain $x \geq 0$ and scale parameter $s > 0$ and shape parameter $\alpha > 0$ (the location is set to 0), has the following density function f (following the notation of Smith and Naylor (1987)):

$$f_{s,\alpha}(x) = \frac{\alpha}{s}\left(\frac{x}{s}\right)^{\alpha-1}\exp\left\{-\left(\frac{x}{s}\right)^{\alpha}\right\}. \tag{2.8}$$

By a similar reasoning as for the Fréchet distribution, I would obtain the likelihood function and hence the derived log-likelihood equations for the numerical estimations of s and α. As for the Fréchet distribution, the estimation of the parameters showed to be difficult, so I used a quantile-regression approach, as explained in section 3, instead. The resulting parameter estimations can be found in table 3.1. What is special about the Weibull distribution is, that this distribution is the only extremal type distribution, that has a right endpoint. If

this distribution can be fitted well to the data, a finite endpoint for the underlying extremal values would be plausible.

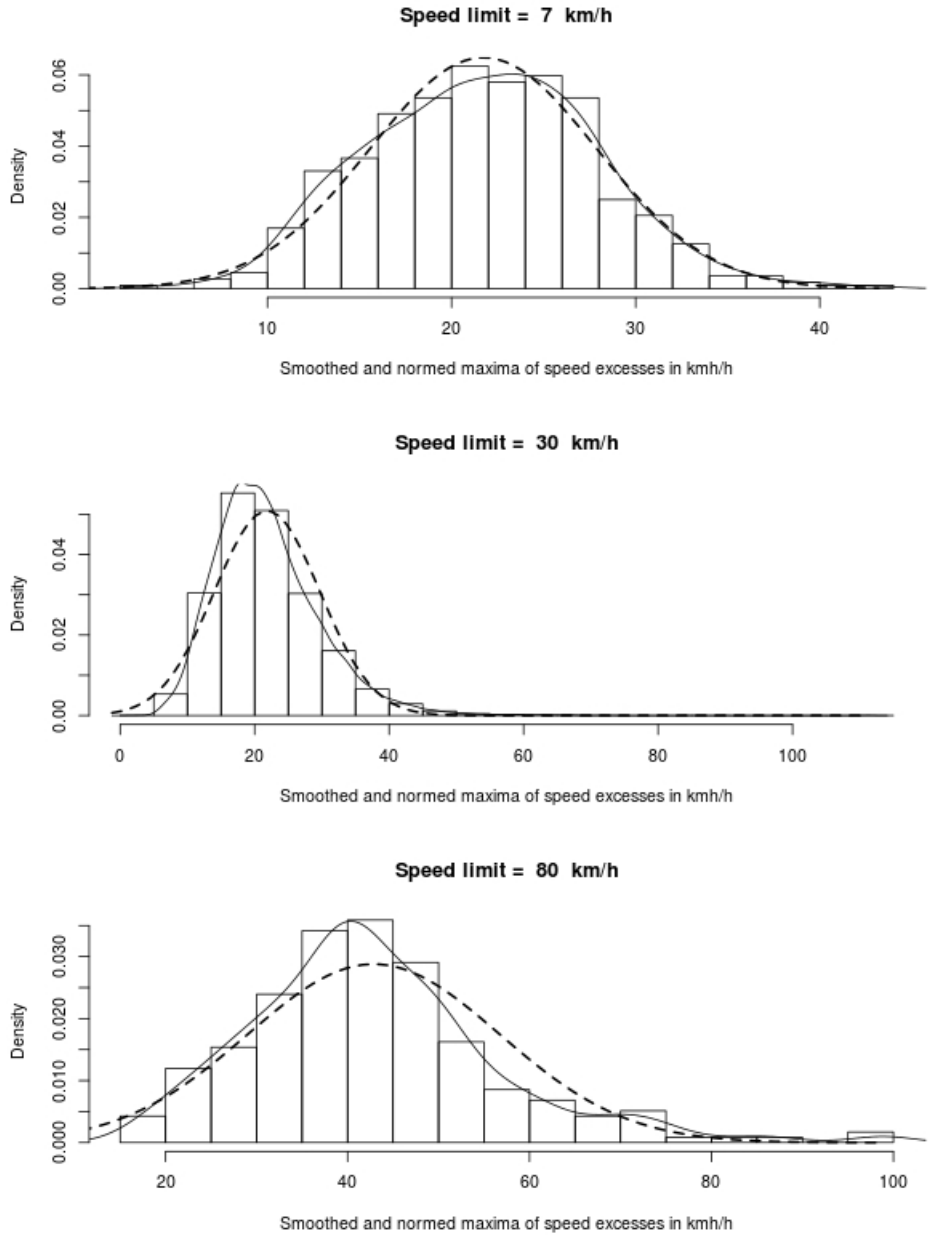

Figure 2.2: Histograms and densities of smoothed and normalized speed excess values in km/h for 7, 30 and 80 km/h. The dashed line represents a density function of a Gaussian distribution with parameters according to table 2.1. The straight line represents a kernel density estimation from the corresponding subset.

2.3.5 Histograms

For the histograms I used the statistical software `R` that calculates the breakpoints by the formula proposed by Sturges (1926). Further information on the theoretical background of histograms as a density estimation and their bias can be found in (Härdle et al., 2004, p. 21-34). Figure 2.2 underlines the idea of using an extreme value approach for the data. The fat tail of the smoothed and normalized maxima of the speed measurements differs from a Gaussian normal distribution. Furthermore, the positive skew of the data becomes evident. The use of distribution functions, that take the negative skew and the left endpoint given by 0 into account, is straightforward. This motivates the use of the three extremal types distributions. These results are evident for all speed classes that are represented in figure 2.2 (see the appendix for the other histograms in figure A.1, A.2 and A.3). As already implied by the box plot in figure 2.1, I expect a detailed analysis by extreme value theory to reveal the heterogeneity within the speed classes. The observations for 7 and 20 km/h, at a first glance, suggest a modeling by a Gaussian distribution, whereas this approach seems less appropriate for 30, 50 and 80 km/h. The pooled data also suggests an approach different to a Gaussian model.

2.3.6 Kernel density estimations

Like the estimation of a histogram, the estimation of a kernel density is a nonparametric method. The kernel density estimation of a set of observations is defined by:

$$\hat{f}_h(x) := \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{2.9}$$

where n denotes the number of observations, h the bandwidth and X_i is the ith observation. I use the Gaussian kernel, given by

$$K(u) = \frac{1}{\sqrt{(2\pi)}} \exp(-\frac{1}{2}u^2).$$

The bias of a kernel density estimation is given by

$$\begin{aligned} \text{Bias}\left\{\hat{f}_h(x)\right\} &= \mathbb{E}\left[\hat{f}_h(x)\right] - f(x) \\ &= \int \frac{1}{h} K\left(\frac{x-u}{h}\right) f(u)du - f(x). \end{aligned}$$

The optimal bandwidth is chosen by $\hat{h} \approx 1.06\hat{\sigma}n^{-1/5}$, where $\hat{\sigma}$ denotes the empirical variance. This bandwidth selector is called Silverman's rule of thumb. See (Härdle et al., 2004,

p. 40-52) for the previous formulas and further details on kernel density estimations.

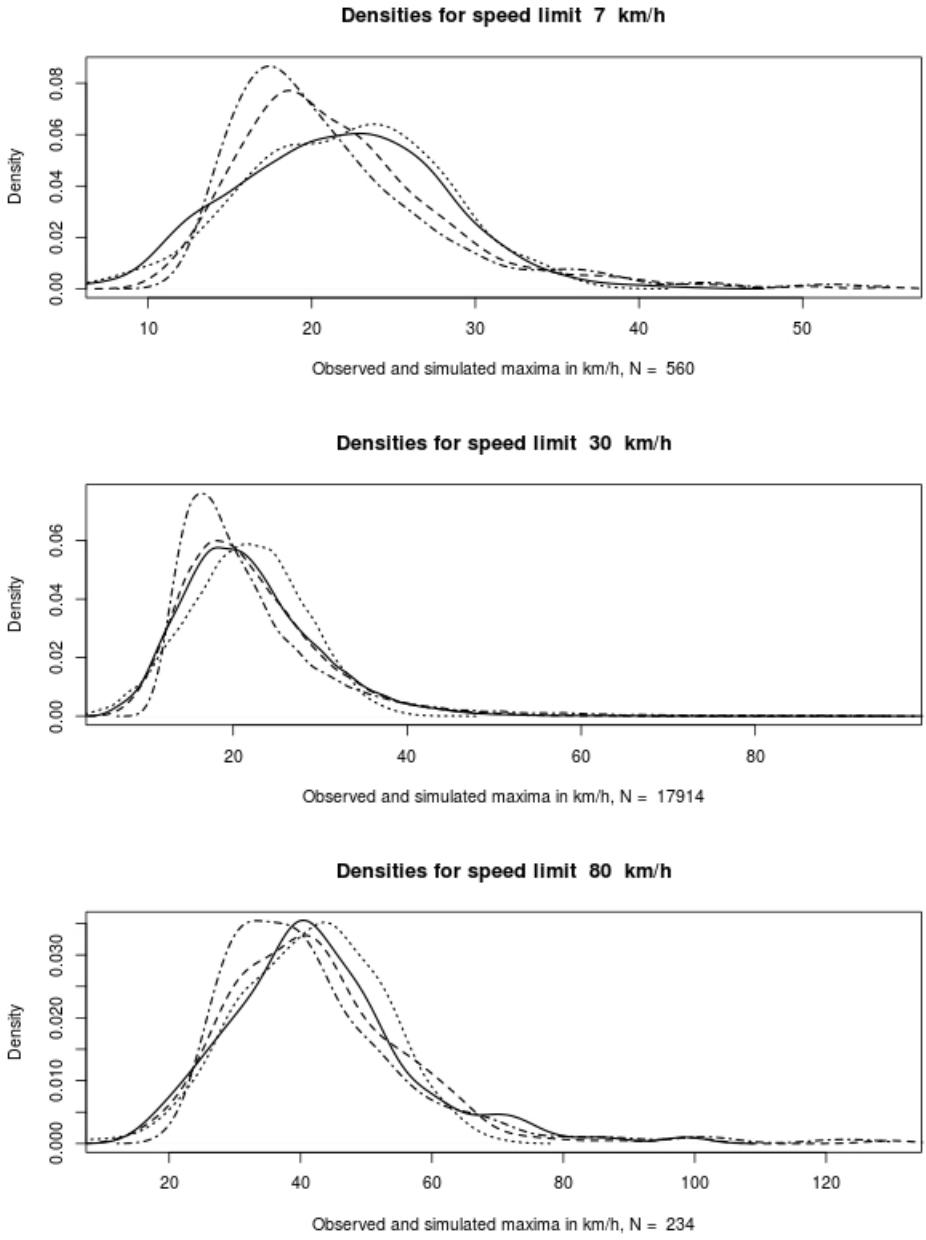

Figure 2.3: Kernel density estimations for the speed limits 7, 30 and 80 km/h of the observed data (straight line), a simulated Weibull distribution (dotted line), a simulated Fréchet distribution (dotted-dashed line) and a simulated Gumbel distribution (dashed line). The parameters used for the simulation can be found in table 2.2 and 3.1. N denotes the number of observations used for the estimation. Observations on the x-axis are given in km/h.

Because of the bias, I cannot compare the kernel density estimations of the observations directly to the density functions of the extreme value distributions. In order to circumvent this problem, I simulate random numbers (using implemented `R`-packages) that follow, for example, a Gumbel distribution with parameters estimated from the data.
Then I can compare the kernel density estimations of this simulated data to the kernel density estimation of the observed values. I apply this procedure to the three extreme value distributions. This yields figure 2.3 (see in the appendix figure A.4, A.5 and A.6 for the other plots). As already presumed, the kernel density estimation of the simulated data which comes closest to the kernel density estimation of the observed data (straight line), is the simulation via a Weibull distribution (dotted line) for 7 and 10 km/h. For 30 and 50 km/h, as well as for the pooled data, I get the impression that a Gumbel distribution (dashed line) could be a good approximation to the real underlying distributions. On the other hand, for 80 km/h it is not clear whether the Fréchet distribution (dotted-dashed line) fits the data. This would mean that there would be no finite endpoint to the speeding data. However, this approach only gives a good approximation for a sufficiently large number of observations. Otherwise, the bias of the kernel density estimations, due to the limited sample for the speed classes, makes it difficult to compare the kernel density estimations. In the next section, I will use a quantile approach to better compare the distributions.

3 Quantile-quantile approach to extreme value theory

3.1 Theoretical idea

To understand whether a statistical model represents a good approximation of a given sample, one often wants to use graphical tools to support or to reject the supposed model. In section 2, I already tried to model the speed maxima for the three extreme value distributions by estimating the parameters of the underlying density functions. The plot of kernel density estimations of the simulated extreme value distributions helped to justify an extreme value approach. Nevertheless, this idea represents a basic approach, that I want to deepen in this section. My aim is not to use inference in order to test the supposed distribution, but to use quantile-quantile-plots, that help understanding the tail-behavior of the extreme values. This approach is widely used in the analysis of extreme values and provides a simple and powerful tool. See (Beirlant et al., 2004, p. 1-4) or (Beirlant et al., 1996, p. 18-37) for further details and applications. The theoretical cumulative distribution function is defined by

$$F(x) = \mathbb{P}(X \leq x), \tag{3.1}$$

and the theoretical quantile function is given by

$$Q(p) := \inf\{x : F(x) \geq p\}. \tag{3.2}$$

The empirical counterpart is given by

$$\hat{F}_n(x) = \frac{i}{n} \text{ if } x \in [x_{i,n}, x_{i+1,n}[\tag{3.3}$$

and

$$\hat{Q}_n(p) := \inf\{x : \hat{F}_n(x) \geq p\}. \tag{3.4}$$

When the theoretical distribution is close to the empirical distribution, I would expect to obtain a straight line when plotting the quantile-quantile tuple given by $\left\{Q(p), \hat{Q}_n(p)\right\}$ for a number of different p. By adding a straight 45° line to the plots, I am then able to easily judge the quality of the approximation by the supposed theoretical distribution. Moreover, it would be possible to fit an ordinary least square (ols) regression to the quantile-quantile-plots to obtain further information on the quality of the model. However, I will not use this method for the Gaussian and Gumbel model. But, I will use this method for the Weibull and Fréchet distribution to judge the parameter estimation from a quantile-plot regression.

3.2 Parameter estimation via quantile-plot regression

By supposing that the data can be modeled by a known distribution function, the question arises whether other methods than moment and maximum-likelihood estimators can be found to determine the parameters of the according distribution functions. The plot of the observed values versus the empirical quantiles represents a possible approach. Given an invertible and left-continuous distribution function, the theoretical quantiles can be calculated from the estimated empirical quantiles. The log-linearization of these equations then gives me a model-equation that I can estimate via ols-regression.

3.2.1 The Fréchet case

After the introduction of the Fréchet distribution in section 2, the cumulative distribution function of a Fréchet-type distribution, with a location parameter = 0 and a scale parameter s and shape parameter α, is given by

$$F\left(X \leq x\right) = \exp\left\{-\left(\frac{x}{s}\right)^{-\alpha}\right\}, \tag{3.5}$$

see (Coles, 2001, p. 46). Let us set $F\left(X \leq x\right) = p_x$, where p_x denotes the probability that a realization X is inferior to a given value x. I obtain p_x by using the empirical quantile-function and plotting these values versus the according values of X from the sample. I then log-linearize this equation:

$$\begin{aligned}
\exp\left\{-\left(\frac{x}{s}\right)^{-\alpha}\right\} &= p_x && (3.6)\\
\left(\frac{x}{s}\right)^{-\alpha} &= -\ln(p_x) && (3.7)\\
-\alpha(\ln(x) - \ln(s)) &= \ln(-\ln(p_x)) && (3.8)\\
\ln(x) &= -\frac{1}{\alpha}\ln(-\ln(p_x)) + \ln(s). && (3.9)
\end{aligned}$$

This results in a linear regression model, where I explain $\ln(x)$ as a linear function of $\ln(-\ln(p_x))$: $\ln(x) = \beta_0 + \beta_1 \ln(-\ln(p_x))$. I get estimates for the coefficients by $\hat{\beta}_0$ and $\hat{\beta}_1$, and by transforming the parameters of the regression, I obtain estimates for $\hat{\alpha} = \frac{-1}{\hat{\beta}_1}$ and $\hat{s} = \exp(\hat{\beta}_0)$. The linearity can easily be checked through the $\{\ln(x), \ln(-\ln(p_x))\}$ plots and the quality of the fit can be quantified by using the results for R^2. This procedure is analogue to the procedure proposed by (Beirlant et al., 1996, p. 28-34) and (Beirlant et al., 2004, p. 3-11).

3.2.2 The Weibull case

For the Weibull distribution, with the probability distribution function

$$F\left(X \leq x\right) = 1 - \exp\left\{-\left(\frac{x}{s}\right)^{\alpha}\right\}, \tag{3.10}$$

with scale parameter s and shape parameter α, I will apply the same method as before. A similar reasoning as for the Fréchet case leads to the following linear function:

$$\ln(x) = \frac{1}{\alpha}\ln(-\ln(1-p_x)) + \ln(s). \tag{3.11}$$

Once more, I will use an ols-estimation to model $\ln(x)$ as a linear function of $\ln(-\ln(1-p_x))$ by $\ln(x) = \beta_0 + \beta_1 \ln(-\ln(1-p_x))$. This will result in the estimations $\hat{\beta}_0$ and $\hat{\beta}_1$. Consequently, I will obtain estimations for the Weibull distribution parameters by $\hat{\alpha} = \frac{1}{\hat{\beta}_1}$ and $\hat{s} = \exp(\hat{\beta}_0)$. See (Beirlant et al., 1996, p. 24-25) and (Beirlant et al., 2004, p. 3-11) for this derivation.

3.3 Results

3.3.1 Fréchet and Weibull quantile-plots

The quantile-regression approach results in the estimations that can be found in table 3.1. For the speed limit 7 km/h the Weibull distribution seems to fit perfectly the observed data, given by a high R^2 of 99.1%. On the other hand, for the speed limit 40 km/h the use of the R^2 is ambiguous, since both the Weibull and the Fréchet distribution provide a R^2 of nearly 93%. In general, the Weibull R^2 is higher than the R^2 of the Fréchet models.

	Fréchet			Weibull			
Speed class in km/h	$\hat{\alpha}$	$\hat{s}$	R^2	$\hat{\alpha}$	$\hat{s}$	R^2	Best fit
7	4.282	18.269	0.866	4.249	23.785	0.991	Weibull
10	4.034	17.303	0.881	4.050	22.855	0.985	Weibull
20	4.702	21.712	0.879	4.729	27.561	0.979	Weibull
30	3.565	17.562	0.932	3.711	23.920	0.970	Weibull
40	2.680	21.408	0.930	2.846	32.155	0.930	Fréchet/Weibull
50	3.172	18.169	0.943	3.336	25.667	0.961	Weibull
60	3.255	24.394	0.896	3.304	34.373	0.980	Weibull
70	3.300	20.242	0.908	3.380	28.344	0.976	Weibull
80	4.039	35.502	0.934	4.229	46.596	0.960	Weibull
overall	3.295	17.999	0.943	3.463	25.104	0.962	Weibull

Table 3.1: Fréchet and Weibull distribution parameters estimated via quantile regression. Numbers are rounded to three decimal points. Estimation errors are omitted.

This is supported by the plots in figure 3.1 or 3.2, which visualize the theoretical idea of the previous section (see the appendix for similar plots of the other speed classes).

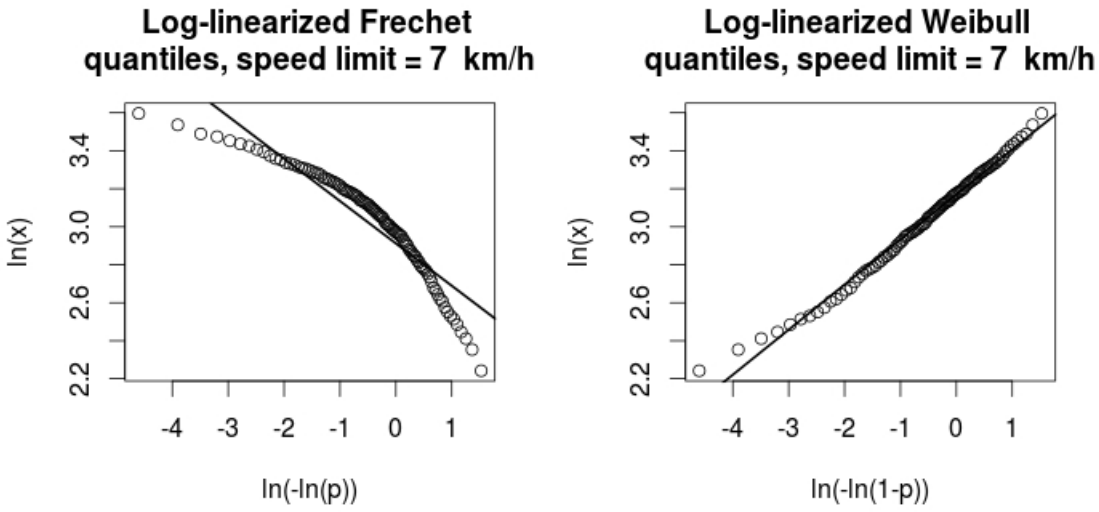

Figure 3.1: Quantile-plots for the log-linearized quantiles by supposing a Fréchet or a Weibull distribution for the speed limit 7 km/h. The straight line represents the linear regression estimated from the data.

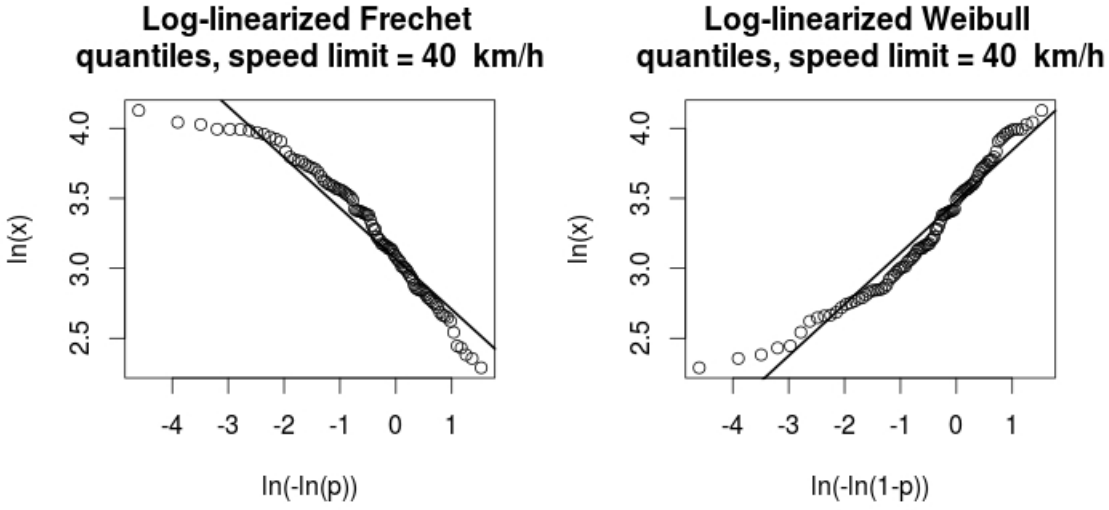

Figure 3.2: Quantile-plots for the log-linearized quantiles by supposing a Fréchet or a Weibull distribution for the speed limit 40 km/h. The straight line represents the linear regression estimated from the data.

The straight line added to the plots represents the linear regression I calculated. The closer

the quantiles are to this straight line, the better the real distribution can be approximated by the supposed distribution. The estimated results are fostered by the graphical analysis. The Weibull distribution seems to fit best the speed limit of 7 km/h. For 40 km/h, both distributions suit well for the speed values. Furthermore, these plots contain additional information on the underlying tail. For 7 km/h, the right tail of the real distribution is fatter than the tail of the Weibull distribution and thinner than the tail of the Fréchet distribution. This could be an indicator that perhaps a Gumbel distribution, with a tail behavior between these two, could even fit the data better. For 40 km/h the tail behavior is similar, whereas the Fréchet distribution fits best for intermediate values of this speed class.

3.3.2 Gaussian and Gumbel quantile-quantile-plots

In the following, I will take a look at quantile-quantile-plots for the Gumbel and Gaussian distributions, whose parameters can be found in table 2.1 and 2.2. The analysis of these figures, where I plot the empirical quantiles versus the theoretical quantiles of a supposed Gaussian and Gumbel distribution, helps to narrow the class of distribution functions for the speed values. If the plot of the empirical and theoretical quantiles are lying on a straight line, this is an indicator that the empirical distribution functions can be approximated well by the theoretical distribution functions.

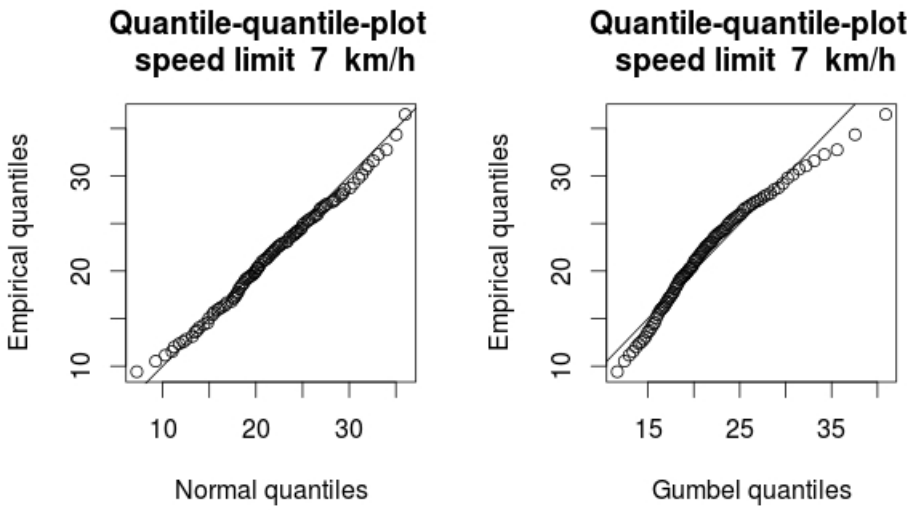

Figure 3.3: Quantile-quantile-plots for the empirical quantiles and the theoretical quantiles of a Gaussian and a Gumbel distribution for the speed limit of 7 km/h. The according parameters for the two distributions are estimated from the data and can be found in table 2.1 and 2.2.

For example, for the speed limit of 7 km/h, as in figure 3.3, I see that the tail of Gumbel distribution is too fat for this speed class. Thus, I might expect to find a right endpoint. I also see that the Gaussian distribution represents a fair model, if I would not take into account the fact that I am merely dealing with maxima of a sample. For the pooled speed data in figure 3.4, the Gumbel distribution represents a good fit. Hence, I do not necessarily expect the existence of a right endpoint for the speeding values. For 10, 20 and 40 km/h, neither a Gaussian, nor a Gumbel distribution do represent a good fit. For 30, 50, 60, 70 and 80 km/h, the Gumbel distributions seem to come close to the empirical distributions. Similar plots for all different speed classes can be found in the appendix.

This graphical approach helped to better understand the behavior of the speeding values. I saw that there are several speed classes, that may have a right endpoint and that there are several for which the existence of an endpoint seems less plausible. In the following section, I will introduce the theoretical concepts, that are needed to estimate the extreme value index, combining all extremal type distributions in one common representation. Consequently, the determination of an endpoint will be derived.

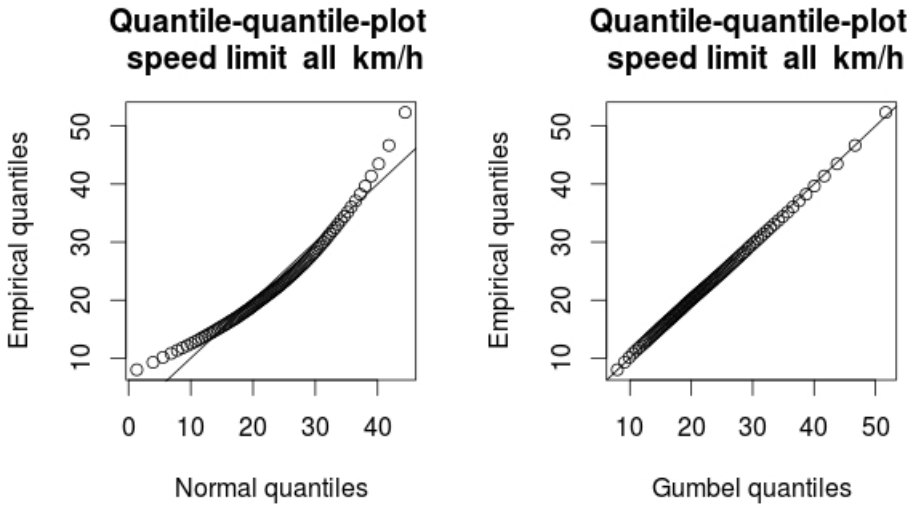

Figure 3.4: Quantile-quantile-plots for the empirical quantiles and the theoretical quantiles of a Gaussian and a Gumbel distribution for the speed limit of 7 km/h. The according parameters for the two distributions are estimated from the data and can be found in table 2.1 and 2.2.

4 Theoretical concepts for extreme value theory

When returning to the question whether the distribution of speeding data has a right endpoint or not, the concepts used so far fail to produce reliable results. I saw that an approximation by the extremal type distributions seemed adequate, but it showed to be difficult to clearly choose one model over the others. Furthermore, the theoretical necessity of this approach was not yet derived. Therefore, the main theoretical concepts of extreme value theory will be explained here. I will introduce several estimation procedures for the extreme value index and how to find the right endpoint. Hence, I will have the necessary concepts at hand to adequately quantify the speeding behavior.

4.1 The special extreme value distributions

For the estimation of the extreme value index, I will first introduce the basic concepts of extreme value theory. For this purpose, I proceed as Einmahl and Magnus did in their paper "Records in Athletics Through Extreme-Value-Theory" (2008).
Let n be the number of maxima observed. Let $X_{1,n}, X_{2,n}, ..., X_{n,n}$ be the upper order statistics of a sequence of independent and identically distributed random variables with the unknown distribution function F. $X_{n,n}$ denotes the absolute maxima. Then it would be natural to choose the following approach for the distribution function:

$$\begin{aligned} \mathbb{P}(X_{n,n} \leq x) &= \mathbb{P}(X_{1,n} \leq x, X_{2,n} \leq x, ..., X_{n,n} \leq x) \\ &= \prod_{i=1}^{n} \mathbb{P}(X_{i,n} \leq x) \\ &= F(x)^n. \end{aligned}$$

However, in practice this is not very helpful, since F is unknown and small deviations in the estimation of F would lead to a large error in the estimation of the distribution function. As explained in (Coles, 2001, p. 45-46), another idea would be to accept that F is unknown and to search for other models for F^n, which will then be estimated on the bases of extreme data only. This is similar to the usual practice of approximating the distribution of sample means by a Gaussian distribution like in the central limit theorem. Let us take a look at the behavior of F^n when $n \to \infty$. For any $x < x_+$, where x_+ denotes the upper endpoint of F such that $x_+ = \inf\{x : F(x) = 1\}$ $\forall x \in \mathbb{R}$ and $\forall x < x_+$, we have $F(x)^n \to 0$ when $n \to \infty$. Therefore, the distribution of $X_{n,n}$ degenerates to a Dirac measurement. This problem can be solved by introducing a linear normalization of $X_{n,n}$ such that $X^*_{n,n} = \frac{X_{n,n} - b_n}{a_n}$, for sequences $\{a_n\}$ $(a_n > 0)$ and $\{b_n\}$. An appropriate choice of theses sequences stabilizes the location

and scale of $X^*_{n,n}$. I therefore seek limit distributions of $X^*_{n,n}$ with adequate choices of $\{a_n\}$ and $\{b_n\}$.

This yields the following extremal type theorem (see (Coles, 2001, p. 46, Theorem 3.1)):

Theorem 1. *If sequences* $\{a_n\}$ $(a_n > 0)$ *and* $\{b_n\}$ *exist such that*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{X_{n,n} - b_n}{a_n} \leq x\right) = G(x), \tag{4.1}$$

where G *is a non degenerate distribution function, then* G *belongs to one of the three following families for* $a > 0$, $b \in \mathbb{R}$ *and* $\alpha > 0$ *:*

$$I: G(x) = \exp\left\{-\exp\left[-\left(\frac{x-b}{a}\right)\right]\right\}, \text{if} -\infty < x < \infty, \tag{4.2}$$

$$II: G(x) = \exp\left\{-\left(\frac{x-b}{a}\right)^{-\alpha}\right\}, \text{if } x > b, \text{otherwise } 0, \tag{4.3}$$

$$III: G(x) = \exp\left\{-\left[-\left(\frac{x-b}{a}\right)\right]^{\alpha}\right\}, \text{if } x < b, \text{otherwise } 1. \tag{4.4}$$

This theorem states that the rescaled sample maxima $\frac{X_{n,n}-b_n}{a_n}$ converge in distribution to a variable having a distribution within one of the families I, II and III. These three classes of distributions are named the extreme value distributions. Distribution I is the Gumbel distribution, II the Fréchet distribution and III the Weibull distribution. Each of the families have a location and scale parameter b and a. Additionally, the Fréchet and Weibull distribution also have a shape-parameter α. The theorem implies that when $X_{n,n}$ can be stabilized with suitable sequences $\{a_n\}$ $(a_n > 0)$ and $\{b_n\}$, the corresponding normalized variable $X^*_{n,n}$ has a limiting distribution that must be one of the three mentioned above. It is quite remarkable that the three distributions are the only possible ones regardless of F. Thus, the theorem provides an analog of the central limit theorem. See (Coles, 2001, p. 45-47).

4.2 The generalized extreme value distribution

The three types of extreme value distributions show different behavior according to the tail behavior of F. This becomes clear by regarding the upper endpoint x_+, which for the Gumbel and Fréchet distribution is infinite and for the Weibull distribution is finite. Moreover, the tail of the Gumbel distribution decays exponentially, whereas the tail of the Fréchet distribution decays with a polynomial degree. Thus, the three types of extreme value distribution represent quite different extreme value behavior. In early applications, like I did in section 2, it was common to estimate the different parameters for the three different type of extreme

value distributions. However, this approach has two weaknesses: firstly, a rule is needed which distribution to choose (I tried to exploit the quantile-quantile approach), and secondly, the use of inference then supposes that the chosen distribution is correct. See (Coles, 2001, p. 47, l. 17-32).

A reformulation of the prior theorem, as in Einmahl and Magnus (2008), combines the three type of distributions in one single distribution, assuming that the values are suitably centered and scaled. This distribution is then called the generalized extreme value distribution (see (Coles, 2001, p. 48, Theorem 3.1.1) and (Einmahl and Magnus, 2008, p. 1384)):

Theorem 2. *If sequences* $\{a_n\}$ $(a_n > 0)$ *and* $\{b_n\}$ *exist, such that*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{X_{n,n} - b_n}{a_n} \leq x\right) = G_\gamma(x), \tag{4.5}$$

for a non-degenerate distribution function G_γ*, then* G_γ *belongs to the following family:*

$$G_\gamma(x) := \exp(-(1+\gamma x)^{-1/\gamma}),$$

defined on the domain $\gamma \in \mathbb{R}$ *with* x*, such that* $1 + \gamma x > 0$*. We have furthermore the convention that* $(1+\gamma x)^{-1/\gamma} = \exp(-x)$ *for* $\gamma = 0$*. If we have the convergence in distribution, we then say that* F *is in the maximum-domain of attraction of* G_γ *and* γ *is called the extreme-value index. For* $\gamma > 0$*, we get the Fréchet-distribution. For* $\gamma = 0$*, we get the Gumbel-distribution and for* $\gamma < 0$*, we obtain the Weibull-distribution.*

My aim in the following will be the estimation of this extreme value index for the speeding data. The latter theorem implies by taking logarithms that

$$\lim_{t\to\infty} t(1 - F(a_t x + b_t)) = -\log G_\gamma(x) = (1+\gamma x)^{-1/\gamma}, G_\gamma(x) > 0,$$

where t runs through $\mathbb{R}^+$ and a_t and b_t are defined by interpolation. We then may take $b_t = U(t)$ with

$$U(t) := \left(\frac{1}{1-F}\right)^{-1}(t) = F^{-1}\left(1 - \frac{1}{t}\right), t > 1, \tag{4.6}$$

where $^{-1}$ denotes the left-continuous inverse and $U(t)$ is the quantile function for the underlying distribution F. These definitions will be helpful for the determination of an endpoint, since the endpoint is nothing else than a high quantile. See (Beirlant et al., 2004, p. 46-51) for a detailed derivation as well as (Einmahl and Magnus, 2008, p. 1384) and (Coles, 2001, p. 46-51). Moreover, it can be shown that theorem 2 is equivalent to the general extremal domain of attraction condition given by

$$\lim_{t\to\infty} \frac{U(tx) - U(t)}{a(t)} = \frac{x^\gamma - 1}{\gamma}, \quad x > 0, \tag{4.7}$$

where a denotes an auxiliary function (see (Beirlant et al., 2004, p. 49). See (Einmahl and Magnus, 2008, p.1384) and (de Haan and Ferreira, 2006, p. 19) for further information on this account. Now, I need to estimate γ, a_t and b_t. The different approaches to determine γ_l, with $l \in \{1, 2, 3, 4\}$ denoting the different estimators, are given in the following subsection.

4.3 Estimators for the extreme value index and the stabilizing sequences

4.3.1 Hill's estimator

As shown in the section for the quantile-quantile approach, I got the impression that for certain speed classes a Fréchet distribution could potentially fit well the speeding values. Thus, I will obtain a positive extreme value index. As a consequence, I will use the best known estimator in this case which is Hill's estimator (first proposed by Hill (1975)).
The estimator is derived as follows: a distribution function F is in the domain of attraction of G_γ for $\gamma > 0$ if and only if

$$\lim_{t\to\infty} \frac{1 - F(tx)}{1 - F(x)} = x^{-1/\gamma}, \quad \gamma > 0. \tag{4.8}$$

This condition is equivalent to:

$$\lim_{t\to\infty} \frac{\int_t^\infty (1 - F(tx)) \frac{dx}{x}}{1 - F(x)} = \gamma. \tag{4.9}$$

Using partial integration yields

$$\int_t^\infty (1 - F(s)) \frac{ds}{s} = \int_t^\infty (\ln(u) - \ln(t)) dF(u). \tag{4.10}$$

By exploiting this result, we get

$$\lim_{t\to\infty} \frac{\int_t^\infty (\ln(u) - \ln(t)) dF(u)}{1 - F(t)} = \gamma. \tag{4.11}$$

To develop an estimator based on this asymptotic result, the parameter t is replaced by the intermediate order statistic $X_{n-k,n}$ and F is replaced by its empirical distribution function $\hat{F}_n$. This results in Hill's estimator $\widehat{\gamma_1}$ defined by

$$\widehat{\gamma_1}(k) := \frac{\int_{X_{n-k,n}}^\infty (\ln(u) - \ln(X_{n-k,n})) d\hat{F}_n(u)}{1 - \hat{F}_n(X_{n-k,n})}, \tag{4.12}$$

or equivalently

$$\widehat{\gamma_1}(k) := \frac{1}{k} \sum_{i=0}^{k-1} \ln(X_{n-i,n}) - \ln(X_{n-k,n}). \tag{4.13}$$

See (de Haan and Ferreira, 2006, p. 69) for this derivation.
For $\gamma > 0$ and $k, n \to \infty, \frac{k}{n} \to 0$ Hill's estimator is consistent (see (Beirlant et al., 2004, p. 104)).

4.3.2 The first moment estimator

Since I am mostly interested in negative values of γ, I will need an estimator that has good properties in this case. The second estimator I will use, is the moment estimator denoted $\widehat{\gamma_2}$ and proposed by Dekkers et al. (1989), which generalizes Hill's estimator for negative values of γ. For $1 \leq k < n$, we define first

$$M_n^{(r)}(k) := \frac{1}{k}\sum_{i=0}^{k-1}\left(\ln\left(X_{n-i,n}\right) - \ln\left(X_{n-k,n}\right)\right)^r, \quad r = 1, 2, \tag{4.14}$$

and then the moment-estimator is given by

$$\widehat{\gamma_2}(k) := M_n^{(1)}(k) + 1 - \frac{1}{2}\left(1 - \frac{\left(M_n^{(1)}(k)\right)^2}{M_n^{(2)}(k)}\right)^{-1}. \tag{4.15}$$

The estimator can be understood by proceeding as follows: for any $i \in \{1, ..., k-1\}$, we have

$$\log(X_{n-i,n}) - \log(X_{n-k,n}) = \log(\hat{U}_n\left(\frac{n+1}{i+1}\right)) - \log(\hat{U}_n\left(\frac{n+1}{k+1}\right)),$$

where $\hat{U}_n$ denotes the empirical counterpart of the quantile function as defined in equation 4.6. Thus, $\log(X_{n-i,n}) - \log(X_{n-k,n})$ can be seen as an estimate of

$$\log(U_n\left(\frac{n+1}{i+1}\right)) - \log(U_n\left(\frac{n+1}{k+1}\right)) = \log(U_n\left(\left(\frac{n+1}{k+1}\right)\left(\frac{k+1}{i+1}\right)\right)) - \log(U_n\left(\frac{n+1}{k+1}\right)).$$

By setting $t = \frac{n+1}{k+1}$ and $x = \frac{k+1}{i+1}$ in the general domain of attraction condition (equivalent to equation 4.7, see (Beirlant et al., 2004, p. 81)), then for any $i \in \{1, ..., k-1\}$ as $n/k \to \infty$

$$\log(X_{n-i,n}) - \log(X_{n-k,n}) \sim \begin{cases} \frac{a\left(\frac{n+1}{k+1}\right)}{U\left(\frac{n+1}{k+1}\right)}\log\left(\frac{k+1}{i+1}\right), & \text{if } \gamma \geq 0, \\ \frac{a\left(\frac{n+1}{k+1}\right)}{U\left(\frac{n+1}{k+1}\right)}\frac{\left(\frac{i+1}{k+1}\right)^{-\gamma}-1}{\gamma}, & \text{if } \gamma < 0. \end{cases} \tag{4.16}$$

For $k \to \infty$, the following limiting results will be used:

$$\frac{1}{k}\sum_{i=1}^{k-1}\log\left(\frac{k+1}{i+1}\right) \to -\int_0^1 \log(u)du = 1,$$

$$\frac{1}{k}\sum_{i=1}^{k-1}\left[\log\left(\frac{k+1}{i+1}\right)\right]^2 \to \int_0^1 \log(u)^2 du = 2,$$

$$\frac{1}{k}\sum_{i=1}^{k-1}\left\{\left(\frac{i+1}{k+1}\right)^{-\gamma} - 1\right\} \to \int_0^1 (u^{-\gamma} - 1)du = \frac{\gamma}{1-\gamma}, \text{ for } \gamma < 0,$$

$$\frac{1}{k}\sum_{i=1}^{k-1}\left\{\left(\frac{i+1}{k+1}\right)^{-\gamma} - 1\right\}^2 \to \int_0^1 (u^{-\gamma} - 1)^2 du = \frac{2\gamma^2}{(1-\gamma)(1-2\gamma)}, \text{ for } \gamma < 0.$$

Therefore, as $k, n \to \infty$ and $k/n \to 0$, we get the convergence in probability

$$\frac{\left(M_n^{(1)}(k)\right)^2}{M_n^{(2)}(k)} \to^P \begin{cases} \frac{1}{2}, & \text{if } \gamma \geq 0, \\ \frac{1-2\gamma}{2(1-\gamma)}, & \text{if } \gamma < 0. \end{cases} \tag{4.17}$$

The estimator is consistent, since

$$M_n^{(1)}(k) \to^P \begin{cases} \gamma, & \text{if } \gamma \geq 0, \\ 0, & \text{if } \gamma < 0. \end{cases}$$

See (Beirlant et al., 2004, p. 142 l. 12 to p. 143 l. 13) for this derivation and (de Haan and Ferreira, 2006, p. 100-109) for a further theoretical insight.

4.3.3 The second moment estimator

The second moment estimator is merely the second part of the first moment estimator. Since it is well known that Hill's estimator $M_n^{(1)}(k)$ is only valid for $\gamma > 0$, the Hill part of the moment estimator is left out and by keeping the definition for $M_n^{(r)}(k)$ this results in

$$\widehat{\gamma_3}(k) := 1 - \frac{1}{2}\left(1 - \frac{\left(M_n^{(1)}(k)\right)^2}{M_n^{(2)}(k)}\right)^{-1}, \tag{4.18}$$

see (Einmahl and Magnus, 2008, p. 1385-1386). By deriving the first moment estimator, this second moment estimator becomes quite intuitive. For positive γ, the second part of the moment estimator converges to 0 in probability, whereas for negative γ, Hill's estimator converges to 0 in probability. It is in this way that the second moment estimator is a consistent estimator for negative values of γ.

4.3.4 The third moment estimator

The third moment estimator has quite a similar structure to the second moment estimator. But this time, we do not take logarithms. This yields

$$N_n^{(r)}(k) := \frac{1}{k}\sum_{i=0}^{k-1}\left(X_{n-i,n} - X_{n-k,n}\right)^r, \quad r = 1, 2. \tag{4.19}$$

And hence, the estimator:

$$\widehat{\gamma_4}(k) := 1 - \frac{1}{2}\left(1 - \frac{\left(N_n^{(1)}(k)\right)^2}{N_n^{(2)}(k)}\right)^{-1}, \tag{4.20}$$

see (Ferreira et al., 2003, p. 411-412) for further information on this estimator.

4.3.5 Estimation of $\hat{a}_l$ and $\hat{b}_l$

As used by (Einmahl and Magnus, 2008, p. 1384), I define the estimators for $a_n(k)$, $b_n(k)$ and $l = 1, 2, 3, 4$ by

$$\hat{a}_l := \hat{a}_{l,n}(k) := \begin{cases} X_{n-k,n} M_n^{(1)}(k)(1 - \hat{\gamma}_l(k)) & \text{if } \hat{\gamma}_l < 0, \\ X_{n-k,n} M_n^{(1)}(k) & \text{otherwise}, \end{cases} \quad \text{and } \hat{b} := \hat{b}_n(k) := X_{n-k,n}, \tag{4.21}$$

where $b_n(k) = U(n/k)$ and $\hat{b}$ is the empirical counterpart. A further theoretical insight is given by (de Haan and Ferreira, 2006, p. 145-147). It is remarkable that all estimators depend on the tail sample fraction of the upper order statistics. This will eventually represent a practical problem in the estimation of the endpoint, as I will show in the next section.

4.3.6 Endpoint estimation

The right endpoint of the distribution is defined as following: $x^* := \sup\{x : F(x) < 1\}$. This endpoint is the ultimate speed of the distribution of speeding values. For the estimation of this endpoint, I need $\gamma < 0$, otherwise $x^* = \infty$. I take up again formula 4.7:

$$\lim_{t\to\infty} \frac{U(tx) - U(t)}{a(t)} = \frac{x^\gamma - 1}{\gamma}, \quad x > 0. \tag{4.22}$$

For large t a heuristic approach is given by

$$U(tx) \approx U(t) + a(t)\frac{x^\gamma - 1}{\gamma}. \tag{4.23}$$

Because $\gamma < 0$, this yields for large x and by setting $t = n/k$,

$$x^* \approx U\left(\frac{n}{k}\right) - a\left(\frac{n}{k}\right)\frac{1}{\gamma}. \tag{4.24}$$

Then x^* is estimated by

$$\hat{x}_l^* := \hat{b} - \frac{\hat{a}_l}{\hat{\gamma}_l}, \quad l = 1, 2, 3, 4, \tag{4.25}$$

when $\hat{\gamma}_l < 0$, otherwise $\hat{x}_l^* := \infty$, see (Einmahl and Magnus, 2008, p. 1384). A detailed explanation can also be found in (Beirlant et al., 2004, p. 156-158).

After the explanation of the theoretical section, I will now turn to the practical estimation of the extreme value index and finally, the endpoint estimations.

5 Results

For the determination of the left endpoint, if it exists, the estimation of the extreme value index γ is crucial. If I obtain an extreme value index smaller than 0, I may suppose that a right endpoint, and thus an absolute limit for the excess speeding, exists. On the other hand, if the extreme value index is positive or 0, no such finite endpoint will be found. As is showed in section 2 and 3, there were some speed classes for which an approximation by a Weibull distribution seemed appropriate and hence, the existence of an endpoint. Whereas for other speed classes, a Gumbel or a Fréchet distribution represented a fair fit and thus, the existence of a finite endpoint may be rejected. It is quite remarkable that depending on the estimator I use for γ, I obtain different results for the existence of an endpoint. I will discuss this idea in the following subsection.

Along with the practical estimation of the extreme value index $\hat{\gamma}_l$, the question arises, how to estimate the stabilizing sequences $\hat{a}_l$ and $\hat{b}_l$ and hence, the endpoint x^*. From the theoretical view, this problem seems clear as discussed in section 4. But the estimation of these parameters depends on k, the number of observation away from the observed absolute maximum of a given class. And since those estimators highly vary according to the range of k, I will use an heuristic approach and explain the limits of this procedure.

In addition to the estimation of the extreme value index and the endpoint for the different speed classes and the pooled data, I will calculate time depending estimators for the extreme value index. I will analyze whether there are seasonal effects for the pooled data by calculating the extreme value indexes by month. Moreover, I am interested in the question whether there is an increase in excess speeding over the half-years. This question will be answered by plotting the resulting estimates for γ as a function of the time.

5.1 Practical estimation of the extreme value index

A test of the existence of γ could be a first step to the estimation of the extreme value index. However, I will not use this procedure but further information on this question can be found in Dietrich et al. (2002) and Drees et al. (2006). As I saw in the theoretical section, the estimators for γ are a function of k (the number of upper order statistics used for the estimation minus 1). The choice of k, and hence the tail sample fraction, on which to base the estimation, is a difficult problem. Therefore, in practice a heuristic approach is used by plotting $\widehat{\gamma}_l$ versus k and by identifying a stable region where the estimations do not fluctuate much. Typically, for small k the estimator has a high variance and the plot is unstable, for

large k the estimator has a bias (see (Einmahl and Magnus, 2008, p. 1385-1386)). I applied this procedure and identified in general a range of 10% to 30% of the upper order statistics in which these fluctuations became sufficiently small. This leads to the following range for the tail sample fraction: $k_{1,j} = \lfloor N_j * 0.1 \rfloor$, for the lowest value of k for a speed class j where N_j denotes the number of observations, and $k_{2,j} = \lfloor N_j * 0.3 \rfloor$, for the highest value of k. I fixed this range for all speed classes to make the estimations comparable. I then calculated the averages of the estimators for $\widehat{\gamma}_l$ and $l = 1, 2, 3, 4$ as defined in the formulas 4.13, 4.15, 4.18 and 4.20. I obtained

$$\overline{\gamma}_{l,j} := \frac{1}{k_{2,j} - k_{1,j}} \sum_{k=k_{1,j}}^{k_{2,j}} \widehat{\gamma}_{l,j}(k), \tag{5.1}$$

as estimator for the speed class j and the estimator l. Furthermore, I calculated an average of the estimators for $l = 2, 3, 4$, since these estimators, in contrast to Hill's estimator, are defined for negative values of γ. This additional estimator is defined as

$$\overline{\gamma}_{2,3,4,j} := \frac{1}{3} \sum_{l=2}^{4} \overline{\gamma}_{l,j}. \tag{5.2}$$

As I will see, this estimator tends to produce far more negative values of γ than $\overline{\gamma}_2$ alone. This is due to the double-weighting of what I called the second moment estimator (see formula 4.18). This term appears twice in the sum. At this point, one might argue that this procedure distorts the estimation. However, I introduced this method due to non-statistical arguments. An infinite endpoint for speeding omits the fact that today there is indeed a maximum velocity for cars given by their construction. The bias to choose negative values of γ over positive ones, thus, may be seen as a heuristic adjustment.

5.1.1 Extreme value index by speed class

By applying the explained procedure, I obtain the ensuing values of $\overline{\gamma}$ that can be found in table 5.1. For Hill's estimator ($\overline{\gamma}_1$) all estimates are larger than 0, which is natural, due to the consistency properties of this estimator. I see that for the moment estimator ($\overline{\gamma}_2$) there are some speed classes like 7, 20 and 40 km/h for which the estimates are clearly negative. Thus, a Weibull distribution with a finite endpoint might represent a good fit. There are other speed classes like 10, 30, 50 and 80 km/h whose values are close to zero and for 60 km/h the extreme value index is clearly positive. Hence, a Gumbel or a Fréchet distribution might be an adequate fitting model. $\overline{\gamma}_3$ can be found through the identity $\overline{\gamma}_3 = \overline{\gamma}_2$-$\overline{\gamma}_1$. Hence, the resulting extreme value indexes are all negative and in this way might predict endpoints that are too small. $\overline{\gamma}_4$ shows a behavior close to the behavior of $\overline{\gamma}_2$. Finally,

$\overline{\gamma}_{2,3,4} = \frac{1}{3}\overline{\gamma}_2 + \frac{1}{3}\overline{\gamma}_3 + \frac{1}{3}\overline{\gamma}_4$ combines all estimators except Hill's estimator and results in almost only negative estimations for the extreme value index.

speed limit km/h	$\overline{\gamma}_1$	$\overline{\gamma}_2$	$\overline{\gamma}_3$	$\overline{\gamma}_4$	$\overline{\gamma}_{2,3,4}$
7	0.11630	-0.09316	-0.20946	-0.10310	-0.13524
10	0.14988	-0.03797	-0.18785	-0.06374	-0.09652
20	0.15211	-0.12293	-0.27505	-0.16104	-0.18634
30	0.18631	0.01451	-0.17180	0.03579	-0.04050
40	0.22242	-0.36131	-0.58373	-0.29779	-0.41428
50	0.22007	0.02647	-0.19359	0.01085	-0.05209
60	0.23157	0.10326	-0.12830	0.04221	0.00572
70	0.22662	0.00777	-0.21885	-0.02509	-0.07873
80	0.19580	-0.03055	-0.22635	-0.07399	-0.11029
all	0.21400	0.06408	-0.14992	0.05559	-0.01008

Table 5.1: Extreme value index estimations for different speed classes. Numbers are rounded to five decimal points. Estimation errors are omitted.

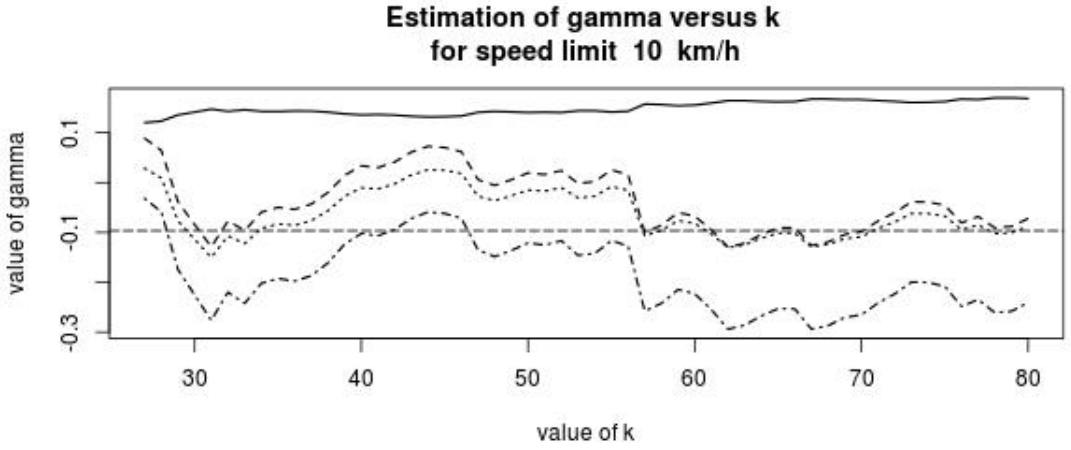

Figure 5.1: Estimation of γ according to the range of the upper order statistics for the speed limit 10 km/h. The straight line represents Hill's estimator $\widehat{\gamma_1}$. The dashed line is the moment estimator $\widehat{\gamma_2}$. The dotted-dashed line represents the second moment estimator $\widehat{\gamma_3}$. And the dotted line represents the third moment estimator $\widehat{\gamma_4}$. The long dashed line is the estimator $\overline{\gamma}_{2,3,4}$. The according estimates can be found in table 5.1.

In figure 5.1, we see the estimates of the extreme value index for the speed limit 10 km/h. There are still fluctuations in the value of $\widehat{\gamma_l}$. Nevertheless, the selection of an even smaller

range of k would reduce the robustness of $\widehat{\gamma_l}$ even more. The resulting extreme value indexes are negative. Consequently, a finite endpoint will be obtained. For the speed limit of 50 km/h, I find that there are hardly any fluctuations in the value of $\widehat{\gamma_l}$. But, the moment estimator $\widehat{\gamma_2}$ tends to produce almost only positive values for γ. As I saw in the previous sections, a Gumbel distribution represented a fair fit for this speed class. So even if I obtain a negative value for $\overline{\gamma}_{2,3,4}$, one should still be aware of the fact that according to the estimation procedure one could influence the resulting endpoint. Similar plots for the other speed limits can be found in the appendix.

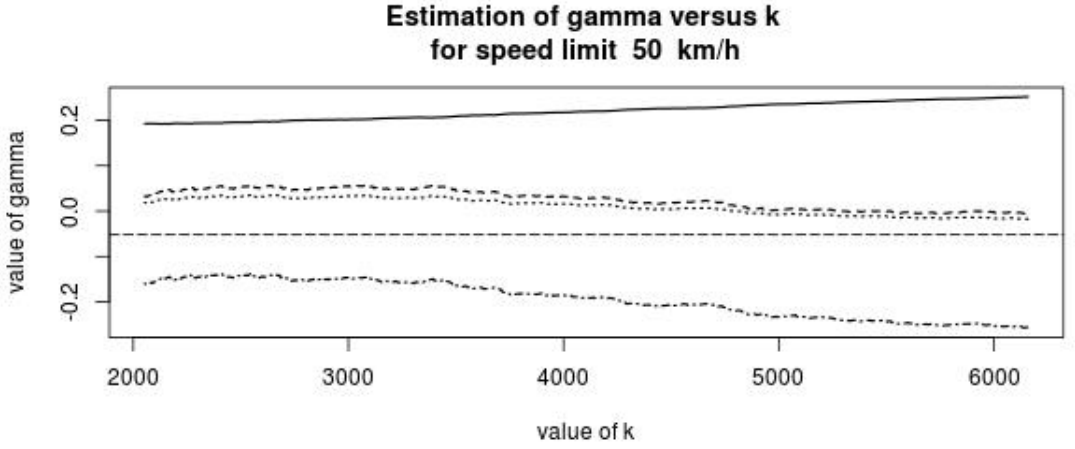

Figure 5.2: Estimation of γ according to the range of the upper order statistics for the speed limit 50 km/h. The straight line represents Hill's estimator $\widehat{\gamma_1}$. The dashed line is the moment estimator $\widehat{\gamma_2}$. The dotted-dashed line represents the second moment estimator $\widehat{\gamma_3}$. And the dotted line represents the third moment estimator $\widehat{\gamma_4}$. The long dashed line is the estimator $\overline{\gamma}_{2,3,4}$. The according estimates can be found in table 5.1.

5.1.2 Extreme value index by month

While analyzing the speeding data, the question arose whether there are seasonal effects in the driving characteristics. For this purpose, I calculated the extreme value index by grouping the pooled data by month. As we see in table 5.2, it will not be possible to determine an endpoint for every month, since some estimates for γ are positive. Nonetheless, the value of the extreme value index itself already indicates a seasonal behavior. This idea is underlined by figure 5.3. A higher extreme value index means a fatter tail of the extreme value distribution and, in this way, higher excess speeding.

Month	$\overline{\gamma}_1$	$\overline{\gamma}_2$	$\overline{\gamma}_3$	$\overline{\gamma}_4$	$\overline{\gamma}_{2,3,4}$
January	0.20662	-0.00006	-0.20668	0.02862	-0.05937
February	0.20647	0.07523	-0.13124	0.05551	-0.00016
March	0.21015	0.03531	-0.17484	0.02439	-0.03838
April	0.21161	0.09553	-0.11607	0.10054	0.02667
May	0.21786	0.08335	-0.13449	0.06307	0.00398
Jun	0.20880	0.01099	-0.19782	-0.01598	-0.06760
July	0.23361	0.06483	-0.16877	0.04575	-0.01939
August	0.22045	0.054027	-0.16642	0.03795	-0.02482
September	0.21395	0.07739	-0.13655	0.06782	0.00289
October	0.21798	0.07737	-0.14061	0.073080	0.00328
November	0.22379	0.04406	-0.17972	0.05205	-0.02787
December	0.18664	0.04104	-0.14559	0.02985	-0.02490

Table 5.2: Extreme value index estimations for different months of the pooled speed limits. Numbers are rounded to five decimal points. Estimation errors are omitted.

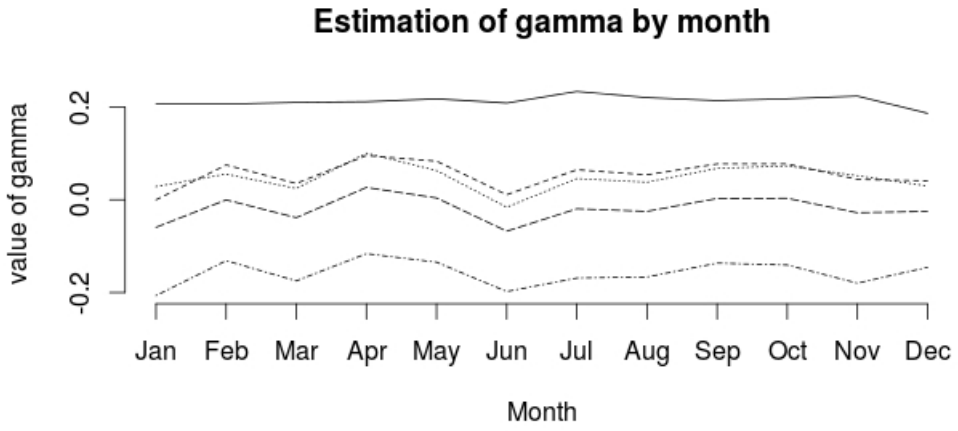

Figure 5.3: Estimation of γ by month for the pooled speed limits. The straight line represents Hill's estimator $\overline{\gamma_1}$. The dashed line is the moment estimator $\overline{\gamma_2}$. The dotted-dashed line represents the second moment estimator $\overline{\gamma_3}$. And the dotted line represents the third moment estimator $\overline{\gamma_4}$. The long dashed line is the estimator $\overline{\gamma}_{2,3,4}$. The according estimates can be found in table 5.2.

By taking into account the estimates for all estimators except Hill's estimator, one could suppose that in June the least extremal speeding can be expected, whereas in April and May, the speeding behavior seems more extreme. In September and October, the extreme value

index rises and then declines in November and January. This spring and fall effect may be explained due to climate conditions. Perhaps, a moderate season leads to more speeding, whereas the winter and summer season reduces excess speeding. Nevertheless, this relation needs further investigation.

5.1.3 Extreme value index by half-year

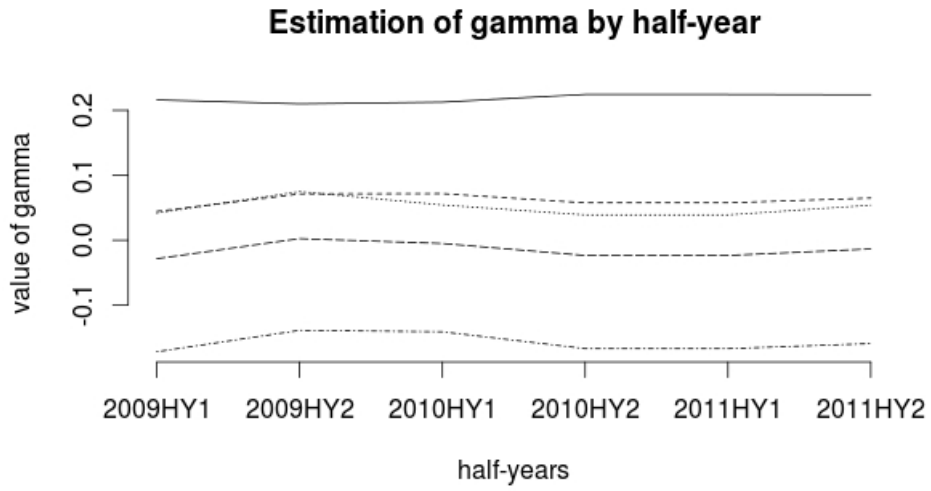

Figure 5.4: Estimation of γ according to the range of the upper order statistics by half-year for the pooled speed limits. The straight line represents Hill's estimator $\overline{\gamma_1}$. The dashed line is the moment estimator $\overline{\gamma_2}$. The dotted-dashed line represents the second moment estimator $\overline{\gamma_3}$. And the dotted line represents the third moment estimator $\overline{\gamma_4}$. The long dashed line is the estimator $\overline{\gamma}_{2,3,4}$. The according estimates can be found in table 5.3.

Half-year and year	$\overline{\gamma}_1$	$\overline{\gamma}_2$	$\overline{\gamma}_3$	$\overline{\gamma}_4$	$\overline{\gamma}_{2,3,4}$
2011HY2	0.22400	0.06499	-0.15902	0.05402	-0.01334
2011HY1	0.20390	0.05495	-0.14895	0.05296	-0.01368
2010HY2	0.22444	0.05767	-0.16678	0.03902	-0.02336
2010HY1	0.21262	0.07178	-0.14084	0.05432	-0.00491
2009HY2	0.20994	0.07129	-0.13865	0.07497	0.00254
2009HY1	0.21607	0.04451	-0.17156	0.04196	-0.02836

Table 5.3: Extreme value index estimations by half-years for the pooled speeding data. Numbers are rounded to five decimal points. Estimation errors are omitted.

In order to find out whether there is a change in speeding behavior over time, I calculated the extreme value index by half-years for the pooled data. The according estimates are given in table 5.3. The plot of the estimated extreme value index by half-year in figure 5.4 does not reveal a significant increase or decrease over time. I tested these effects by a simple ols-regression and for none of the estimators the slope coefficient was significant on a level of significance of 5%. The existence of seasonal effects could not be reinforced. However, I exploited only six observations, what makes it difficult to draw reliable conclusions. The use of time series analysis and a seasonal adjustment could be helpful for further investigation.

5.2 Endpoint estimations

For the endpoint estimation, I use the method explained in section 4, in particular in formula 4.21 and 4.25. Since the stabilizing sequences are a function of k, and in particular $\hat{a}$ is a function of γ, the endpoint estimation is susceptible to variations in γ that might arise due to the estimation procedure. Hence, I fix γ to the values I estimated in table 5.1. This reduces the variability, see (Einmahl and Magnus, 2008, p. 1386). Combining the different equations and supposing that $\gamma < 0$, I obtain finally the estimator for x^* by

$$\hat{x}^*_{j,l}(k) = X_{j,n-k,n} - X_{j,n-k,n} M^{(1)}_{j,n}(k) \frac{(1 - \overline{\gamma}_{j,l})}{\overline{\gamma}_{j,l}}. \tag{5.3}$$

Here once more, I distinguish by speed limit j and extreme value index estimator l. $M^{(1)}_n(k)$ denotes Hill's Estimator as defined in formula 4.13. By following the path I explained for the extreme value index, once again I use the range for the upper order statistics fixed by $k_{1,j} = \lfloor N_j * 0.1 \rfloor$, for the lowest value of k for a speed class j, where N_j denotes the number of observations. And $k_{2,j} = \lfloor N_j * 0.3 \rfloor$ for the highest value of k. This is illustrated by figure 5.5. I plotted the estimations according to formula 5.3 where $\overline{\gamma}_{j,l}$ stays fixed. Then I calculated the mean of all estimates depending on k for the given range. The according results can be found in table 5.4.

The fact that the estimations of a and b is not necessarily stable has a strong influence on the estimation of the endpoint x^*, as shown in figure 5.5 for the speed limit of 40 km/h. Depending on the estimator, the resulting endpoint moves in a bandwidth of more than 15 km/h. But still the resulting endpoints are relatively close to another. This is not the case for the speed limit of 80 km/h.

As figure 5.6 underlines, there exists a high variation in the resulting endpoint estimations. For this speed limit the moment estimator $\overline{\gamma}_2$ suggests an endpoint up to 400 km/h. In this case, it would be more intuitive to prefer the estimations proposed by $\overline{\gamma_3}$ or $\overline{\gamma}_{2,3,4}$, which are

closer to the observed maximum. Similar plots for the other speed classes can be found in the appendix.

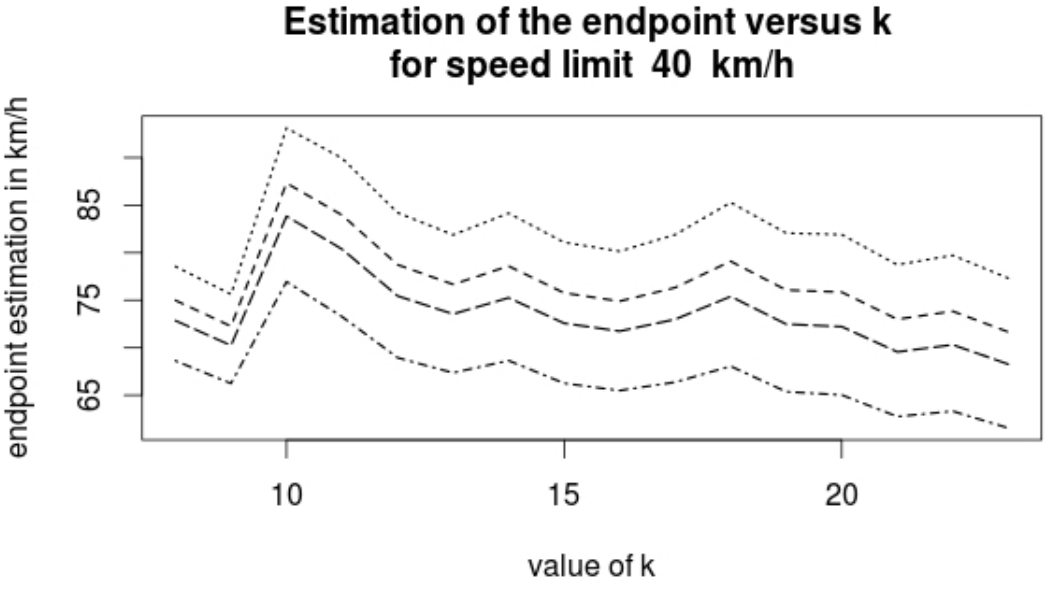

Figure 5.5: Endpoint estimation in km/h of x^* according to the range of the upper order statistics for the speed class of 40 km/h. The dashed line represents the endpoint estimations using the moment estimator $\overline{\gamma_2}$. The dotted-dashed line represents endpoint estimations for the second moment estimator $\overline{\gamma_3}$. And the dotted line represents the endpoints for the third moment estimator $\overline{\gamma_4}$. The long dashed line is the endpoint obtained by the estimator $\overline{\gamma}_{2,3,4}$. The according estimates can be found in table 5.4.

An essential part of the estimated endpoints is quite reasonable regarding the observed absolute maxima. Though, there are some speed classes for which the estimation by a specific estimator failed (as shown in table 5.4 by missing values). This might be an indicator that there is no finite endpoint for this speed limit. Furthermore, there are some endpoints that are implausible since a lower endpoint than the observed absolute maximum is predicted. In this case, the resulting estimations for γ are too negative, which might be due to the preference of estimators that tend to be negative (in particular $\overline{\gamma}_3$ and $\overline{\gamma}_{2,3,4}$).
This idea is taken up by figure 5.7 that visualizes the predicted endpoints and the measured maxima. For the speed limit of 40 km/h, almost no increase in the speeding behavior can be expected. Some of the predicted endpoints are even below the measured maximum. For 7 and 20 km/h a slight increase is still possible. For 10, 70 and 80 km/h as allowed speed, a

further increase is likely. And regarding the pooled speed excesses, one should expect even more aggressive speeding behavior. For 30 and 50 km/h there is also a possible increase in the speeding behavior.

At this point, the limits of the different estimators, as well as the method for the selection of the tail sample fraction, become apparent. $\overline{\gamma}_3$ has tendency to underestimate the extreme value index, since the predicted endpoints are often below the realized maxima. $\overline{\gamma}_2$ predicts finite endpoints merely for the speed limit of 7, 10, 20 and 70 km/h. $\overline{\gamma}_{2,3,4}$ and $\overline{\gamma}_4$ have a similar behavior but predict endpoints up to ≈ 340 and ≈ 660 km/h. Given today's natural constraints, these endpoints are not fully satisfying. In addition to that, the selection of the tail sample fraction influences the results. Pooling the data and fixing the tail sample fraction, leads to endpoint estimations that are far above the endpoints for the other speed classes. This behavior might be explained by a distortion caused by the speed classes of 30 and 50 km/h, since the greatest number of observations are available for these speed limits.

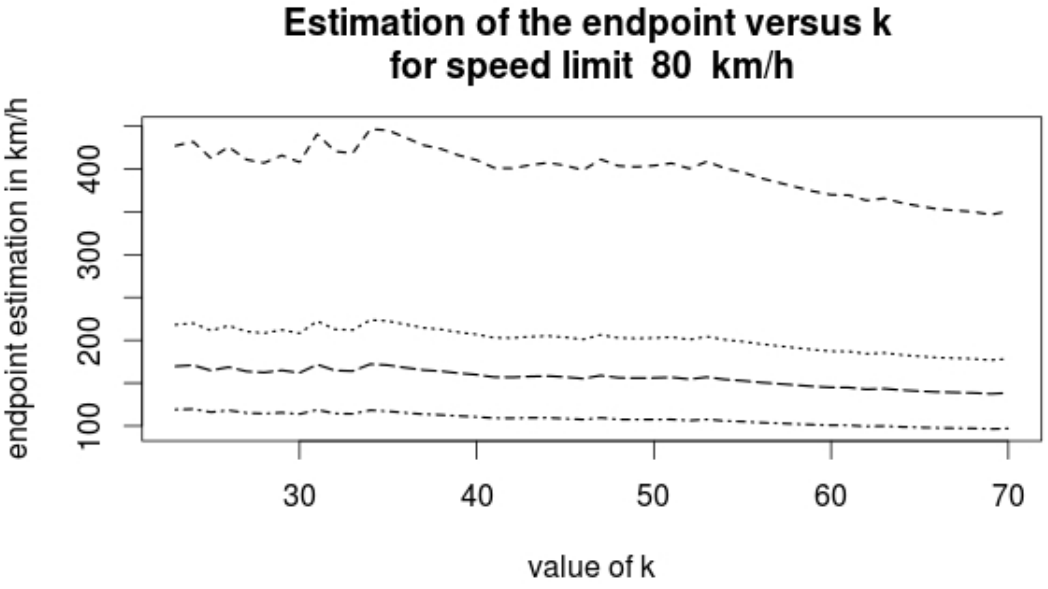

Figure 5.6: Endpoint estimation in km/h of x^* according to the range of the upper order statistics for the speed class of 80 km/h. The dashed line represents the endpoint estimations using the moment estimator $\overline{\gamma_2}$. The dotted-dashed line represents endpoint estimations for the second moment estimator $\overline{\gamma_3}$. And the dotted line represents the endpoints for the third moment estimator $\overline{\gamma_4}$. The long dashed line is the endpoint obtained by the estimator $\overline{\gamma}_{2,3,4}$. The according estimates can be found in table 5.4.

speed limit km/h	$\overline{\gamma}_2$	$\overline{\gamma}_3$	$\overline{\gamma}_4$	$\overline{\gamma}_{2,3,4}$
7	63.63	45.00	60.39	53.19
10	131.14	50.19	90.13	69.60
20	72.64	51.88	63.76	59.86
30	-	63.03*	-	160.34
40	76.82*	67.15*	82.23	73.57*
50	-	71.52*	-	164.95
60	-	119.07	-	-
70	-	75.10	340.53	136.29
80	398.89	108.13	201.53	155.85
all	-	78.04*	-	661.60

Table 5.4: Endpoint estimations for different speed classes. Numbers are rounded to five decimal points. Values denoted by * are implausible, since the observed maximum for the given speed class is higher than the estimated endpoint. Estimation errors are omitted.

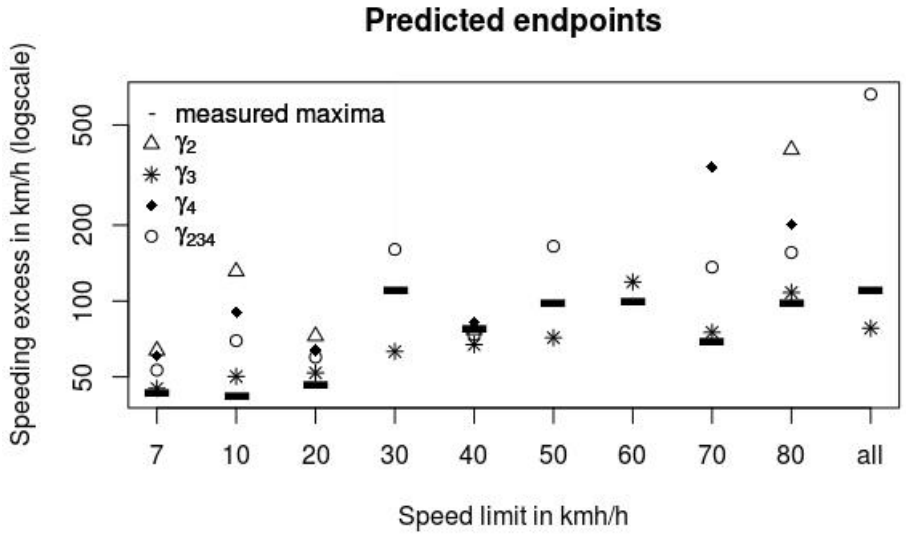

Figure 5.7: Predicted endpoints and maximum measured speed excesses for different speed classes and estimators. The black bars represent the absolute maximum speed excesses that have been measured. The triangles represent the endpoint estimations using the moment estimator $\overline{\gamma_2}$. The stars represent endpoint estimations for the second moment estimator $\overline{\gamma_3}$. The black diamonds represent the endpoints for the third moment estimator $\overline{\gamma_4}$. The dots are the endpoints obtained by estimator the $\overline{\gamma}_{2,3,4}$. The according estimates can be found in table 5.4.

6 Conclusion

Is extreme value theory a reasonable approach to the analysis of speeding data? To some extent my answer is yes. I approached this question by applying basic statistical tools in section 2 where I showed that the available speeding data had fat tails and that a Gaussian modeling approach was not satisfying. Furthermore, the fact that the speeding data was given in a sample of maxima for different dates supported the idea of modeling the sample distributions by the three extremal type distributions. For this purpose, I had to deal with the estimation of the distribution parameters which proved to be a non-trivial question.
While moment and maximum likelihood estimators failed for the Weibull and Fréchet distributions, the question arose what other estimators I had at my disposal. An adequate solution was the parameter estimation by quantile-plot regression to which I turned to in section 3. Exploiting these estimations, I returned to the simulation of the three extreme value distributions using the estimated parameters. I then fitted kernel density estimations to these samples. This method showed that Gumbel and Weibull distributions represented a fair fit for the samples, which supported the idea of using an extreme value approach. Consequently in section 3, I wanted to validate this idea by an even more powerful graphical tool: the use of quantile-quantile plots.
These plots supported the existence of a finite endpoint for several speed classes, whereas for others no finite endpoint seemed plausible. Then in section 4, I derived the theoretical ideas that stand behind extreme value theory, the estimation of the extreme value index and the endpoint estimation. These estimators are then used in section 5 to calculate the extreme value index for the different speed classes.
On the whole, I used five different estimators for the extreme value indexes. Although the properties of the different estimators are clearly defined, the choice of the tail sample fraction had a strong influence on the resulting estimates. In this case, I had to deal with a clear trade-off in choosing a stable region of $\widehat{\gamma}_l(k)$ for the estimation of $\overline{\gamma}_l$, and the use of as much information of the tail sample fraction as possible. To avoid this dependence, I chose a heuristic approach by fixing the tail sample fraction to a range of the upper order statistics. By the same means, I calculated extreme value indexes, on one hand, for the different speed classes and the pooled data and, on the other hand, for the pooled data grouped by half-years and months.
For all speed classes, I found at least one negative estimation for the value of γ. This speaks in favor of the existence of an endpoint for every speed class. Furthermore, I found evidence

of seasonal effects by plotting the estimated extreme value index by month. In spring and fall the driving behavior seemed to be more extreme than in summer and winter. However, an increase or decrease in the driving behavior over time could not be confirmed.

Finally, I calculated the endpoints for the different speed classes according to estimations of γ found before. For this purpose, I had to calculate the stabilizing sequences a and b that depended on γ and k. To avoid strong fluctuations in the resulting endpoint, I fixed γ such that the endpoint became only susceptible to fluctuations in a and b. I used once more the heuristic approach to choose the tail sample fraction.

I saw that depending on the chosen estimator $\overline{\gamma}_l$, the resulting endpoint varied. By the choice of the estimator, one can influence, for example, if the endpoint will be close to the observed maximum or if the resulting γ is so small that one should not assume a finite endpoint. All in all, the calculated endpoints still revealed one fact, given todays vehicle construction: there is still way for excessive speeding behavior, far more extreme than observed until now.

As we saw, the use of the different estimators had an important influence of the resulting values of γ. During this thesis, I covered only estimators for the extreme value index related to Hill's estimator, but there is still a wide variety of other extreme value index estimators that one might use. On one hand, there is the maximum likelihood estimator as proposed by Smith (1987) which might give interesting results. On the other hand, I could have used Pickands's estimator as proposed by Pickands III (1975) and detailed in (Embrechts et al., 1997, p. 327-330).

Moreover, the use of quantile plot regressions to determine the extreme value index could yield new results. In this case, it could be helpful to determine the regression errors to judge its quality. In addition to the employed heuristic approach for the tail sample fraction selection, it could be rewarding to exploit more sophisticated methods for the selection of the values of k, for example, a bootstrap approach as proposed by Danielsson et al. (2001).

The test of the existence of γ as in Dietrich et al. (2002) and Drees et al. (2006) could yield additional information. Analyzing the estimations of γ, using the normality of γ as in Dekkers et al. (1989), could be interesting, too. Furthermore, one could use inference to test the value of x^* and plot the resulting confidence intervals for a given confidence level, since x^* is asymptotically normal (see (de Haan and Ferreira, 2006, p. 147-148) and (Einmahl and Magnus, 2008, p. 1384)).

Besides the improvement of the methods already used, it would be interesting to exploit other information of the dataset. Is there a dependence of measurement period and observed speed

maxima? Or is there a connection between the number of cars recorded, the number of fees imposed and the bans on driving? How do other events like heavy weather conditions or Easter and Christmas holidays affect the driving behavior? And do daytime or rush hours influence the driving behavior? Many questions remain to be answered by future research.

References

Beirlant, J., Y. Goegebeur, J. Segers, J. Teugels, D. De Waal, and C. Ferro (2004): *Statistics of Extremes: Theory and Applications*, Wiley Series in Probability and Statistics, West Sussex: Wiley.

Beirlant, J., J. L. Teugels, and P. Vynckier (1996): *Practical analysis of extreme values*, vol. 50, Leuven: Leuven University Press.

Bertsekas, D. P. and J. N. Tsitsiklis (2002): *Introduction to probability*, vol. 1, Athena Scientific Belmont, MA.

Coles, S. (2001): *An Introduction to Statistical Modeling of Extreme Values*, Lecture Notes in Control and Information Sciences, London: Springer.

Danielsson, J., L. de Haan, L. Peng, and C. G. de Vries (2001): "Using a bootstrap method to choose the sample fraction in tail index estimation," *Journal of Multivariate analysis*, 76, 226–248.

de Haan, L. and A. Ferreira (2006): *Extreme Value Theory: An Introduction*, Springer Series in Operations Research and Financial Engineering, New York: Springer.

Dekkers, A. L., J. H. Einmahl, and L. De Haan (1989): "A moment estimator for the index of an extreme-value distribution," *The Annals of Statistics*, 17, 1833–1855.

Dietrich, D., L. Haan, and J. Hsler (2002): "Testing Extreme Value Conditions," *Extremes*, 5, 71–85.

Drees, H., L. de Haan, and D. Li (2006): "Approximations to the tail empirical distribution function with application to testing extreme value conditions," *Journal of Statistical Planning and Inference*, 136, 3498–3538.

Einmahl, J. H. J. and J. R. Magnus (2008): "Records in Athletics Through Extreme-Value Theory," *Journal of the American Statistical Association*, 103, 1382–1391.

Embrechts, P., C. Klüppelberg, and T. Mikosch (1997): *Modelling Extremal Events: For Insurance and Finance*, Applications of mathematics, 33, Berlin, Heidelberg, New York: Springer.

Embrechts, P., S. I. Resnick, and G. Samorodnitsky (1999): "Extreme value theory as a risk management tool," *North American Actuarial Journal*, 3, 30–41.

Ferreira, A., L. de Haan, and L. Peng (2003): "On optimising the estimation of high quantiles of a probability distribution," *Statistics*, 37, 401–434.

Gumbel, E. J. (1965): "A Quick Estimation of the Parameters in Fréchet's Distribution," *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, 33, 349–363.

Härdle, W., M. Müller, S. Sperlich, and A. Werwatz (2004): *Nonparametric and semiparametric models*, Springer Series in Statistics, Berlin: Springer Verlag.

Hill, B. M. (1975): "A simple general approach to inference about the tail of a distribution," *The annals of statistics*, 3, 1163–1174.

Landwehr, J. M., N. Matalas, and J. Wallis (1979): "Probability weighted moments compared with some traditional techniques in estimating Gumbel parameters and quantiles," *Water Resources Research*, 15, 1055–1064.

Mubarak, M. (2011): "Estimation of the Fréchet Distribution Parameters Based on Record Values," *Arabian Journal for Science and Engineering*, 36, 1597–1606.

Palutikof, J. P., B. B. Brabson, D. H. Lister, and S. T. Adcock (1999): "A review of methods to calculate extreme wind speeds," *Meteorological Applications*, 6, 119–132.

Pickands III, J. (1975): "Statistical inference using extreme order statistics," *the Annals of Statistics*, 3, 119–131.

Smith, R. L. (1987): "Estimating tails of probability distributions," *The annals of Statistics*, 15, 1174–1207.

Smith, R. L. and J. C. Naylor (1987): "A Comparison of Maximum Likelihood and Bayesian Estimators for the Three- Parameter Weibull Distribution," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 36, 358–369.

Sturges, H. A. (1926): "The choice of a class interval," *Journal of the American Statistical Association*, 21, 65–66.

Yee, T. W. (2013): *VGAM: Vector Generalized Linear and Additive Models*, r package version 0.9-1.

0

A Figures

Figures I refer to and not shown in the main part can be found on the following pages.

A.1 Histograms

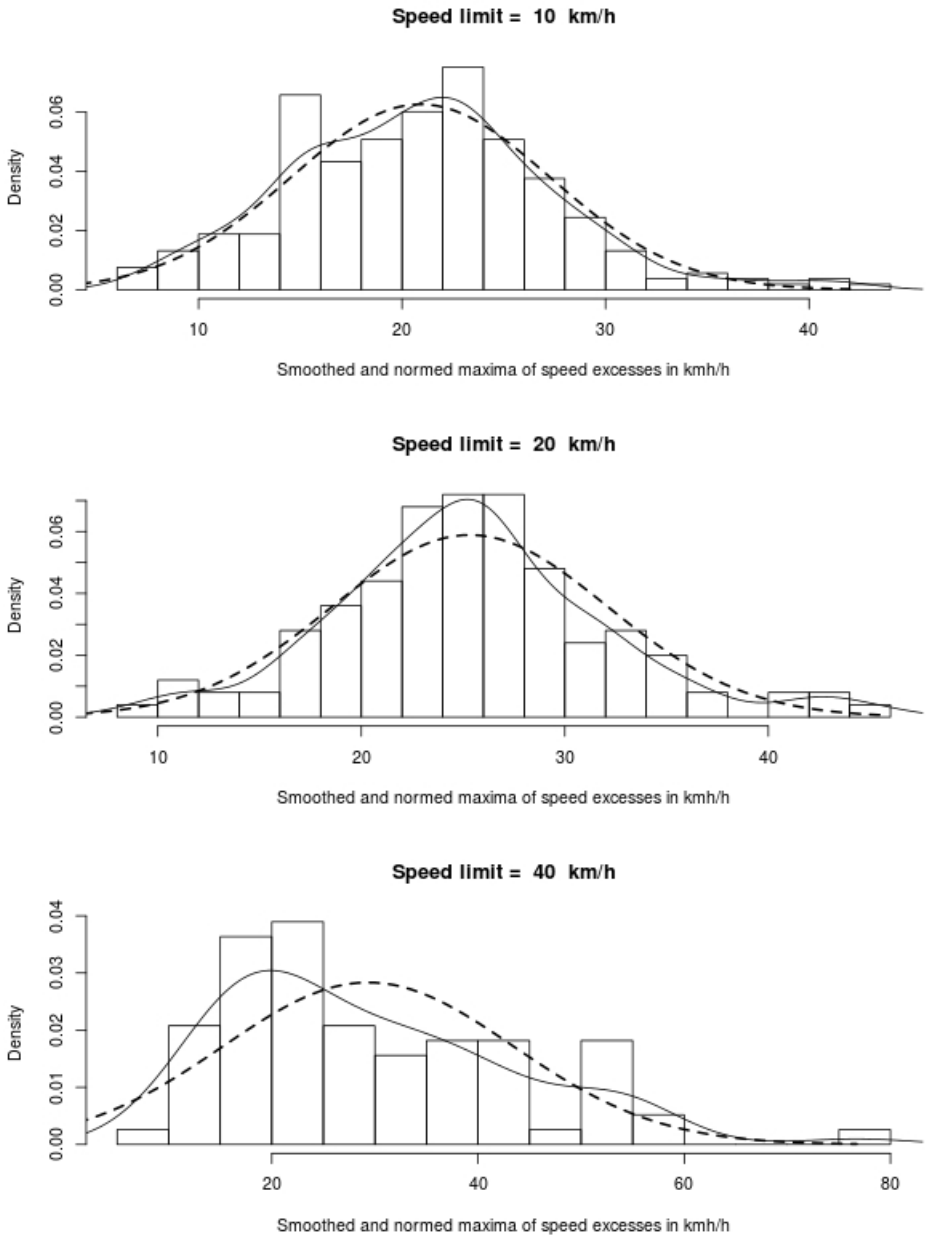

Figure A.1: Histograms and densities of smoothed and normalized speed excess values in km/h for the speed limit 10, 20 and 40 km/h. The dashed line represents the density function of a Gaussian distribution with parameters according to table 2.1. The straight line represents a kernel density estimation from the corresponding subset.

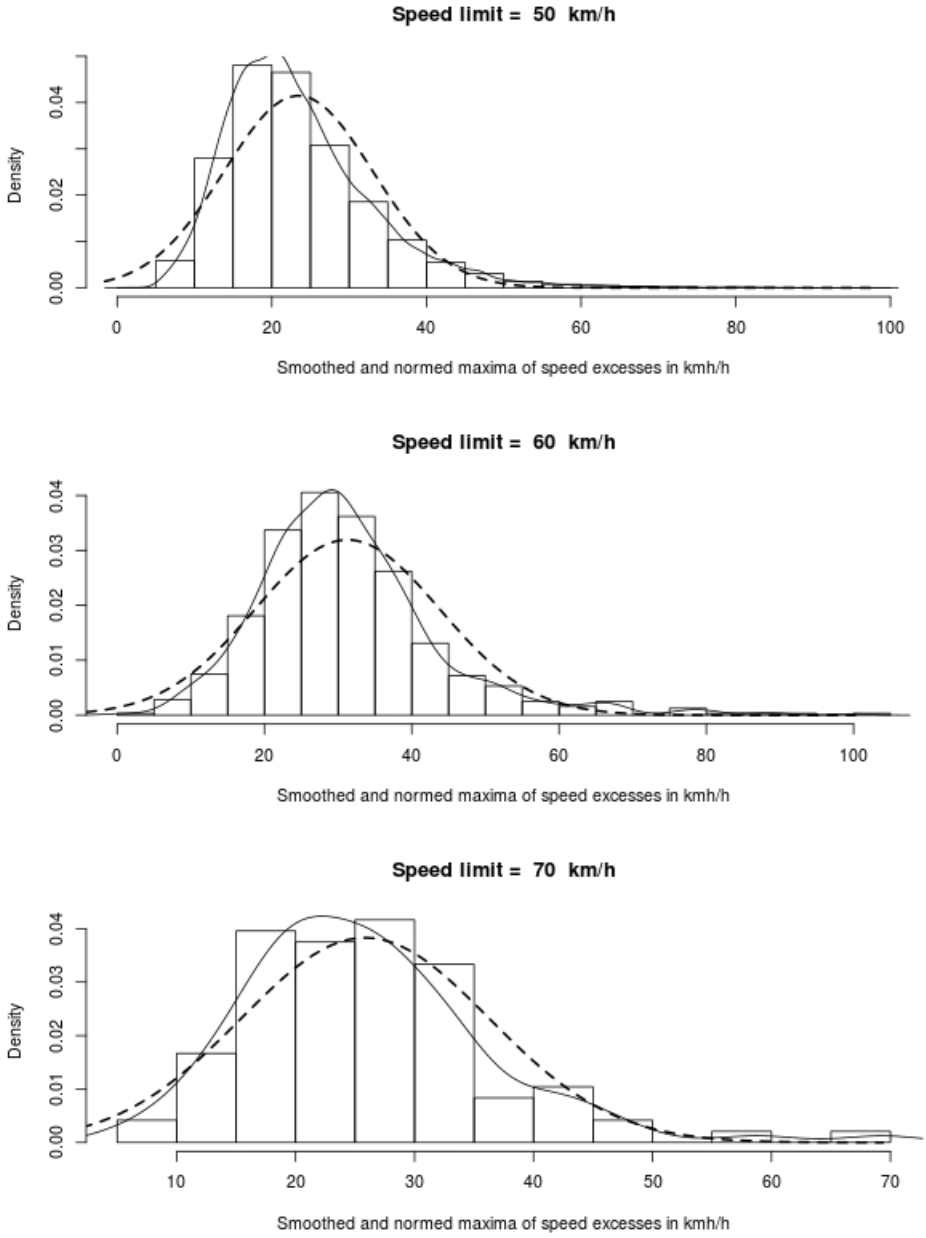

Figure A.2: Histograms and densities of smoothed and normalized speed excess values in km/h for the speed limit 50, 60 and 70 km/h. The dashed line represents the density function of a Gaussian distribution with parameters according to table 2.1. The straight line represents a kernel density estimation from the corresponding subset.

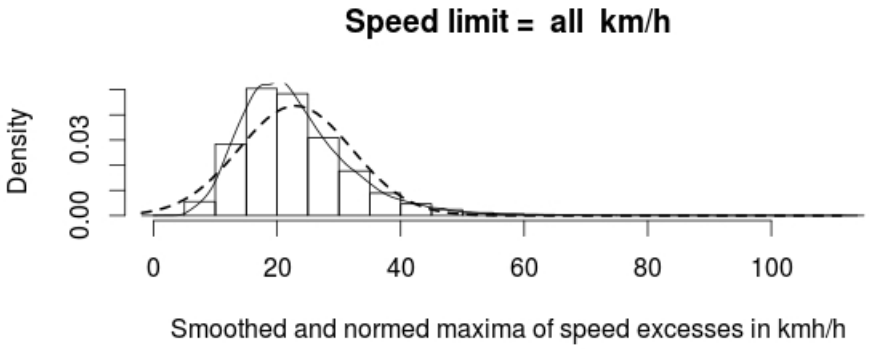

Figure A.3: Histogram and densities of smoothed and normalized speed excess values in km/h for the pooled speeding data. The dashed line represents the density function of a Gaussian distribution with parameters according to table 2.1. The straight line represents a kernel density estimation from the corresponding subset.

A.2 Kernel density estimations

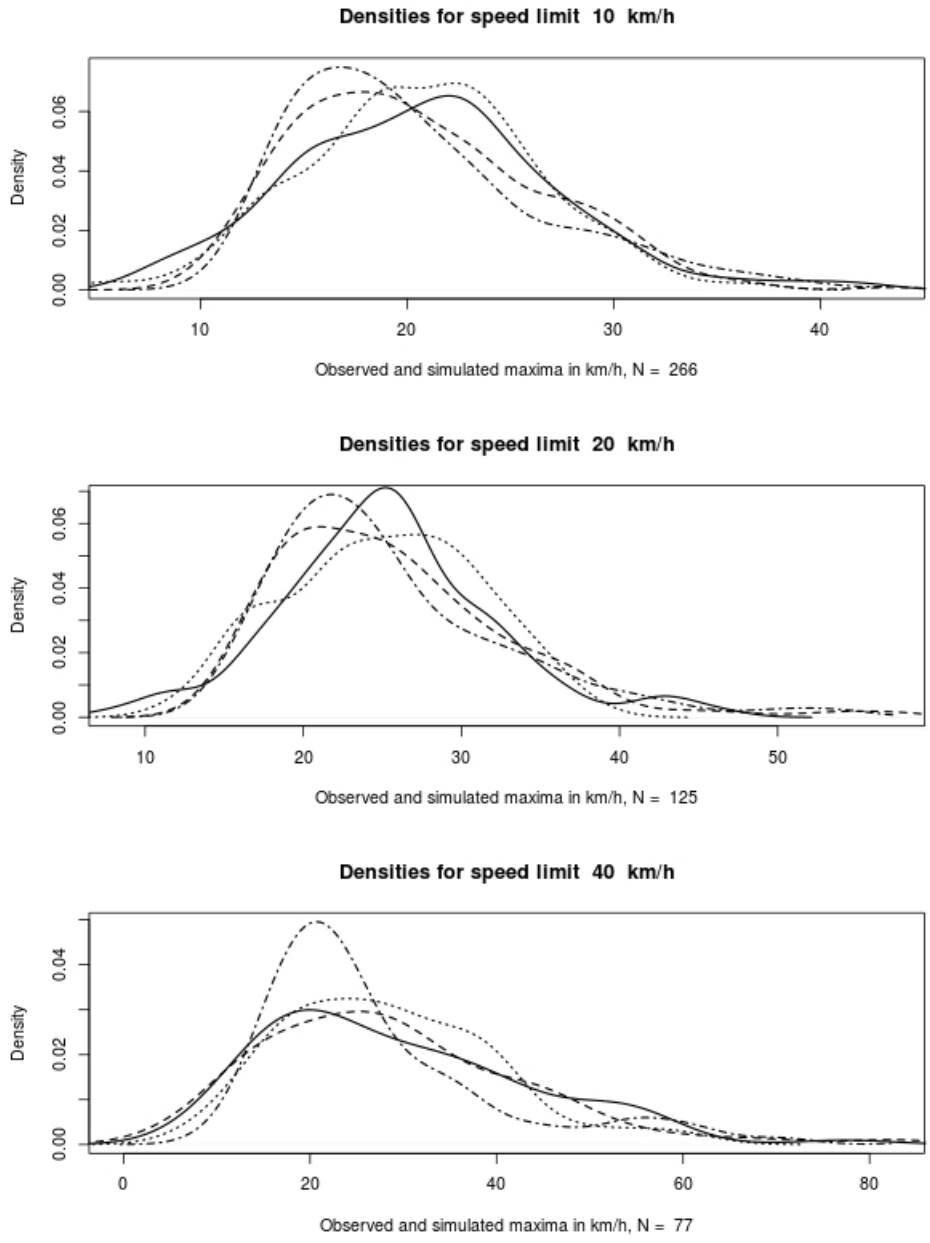

Figure A.4: Kernel density estimations for the speed limits 10, 20 and 40 km/h of the observed data (straight line), a simulated Weibull distribution (dotted line), a simulated Fréchet distribution (dotted-dashed line) and a simulated Gumbel distribution (dashed line). The parameters used for the simulation can be found in table 2.2 and 3.1. N denotes the number of observations used for the estimation. Observations on the x-axis are given in km/h.

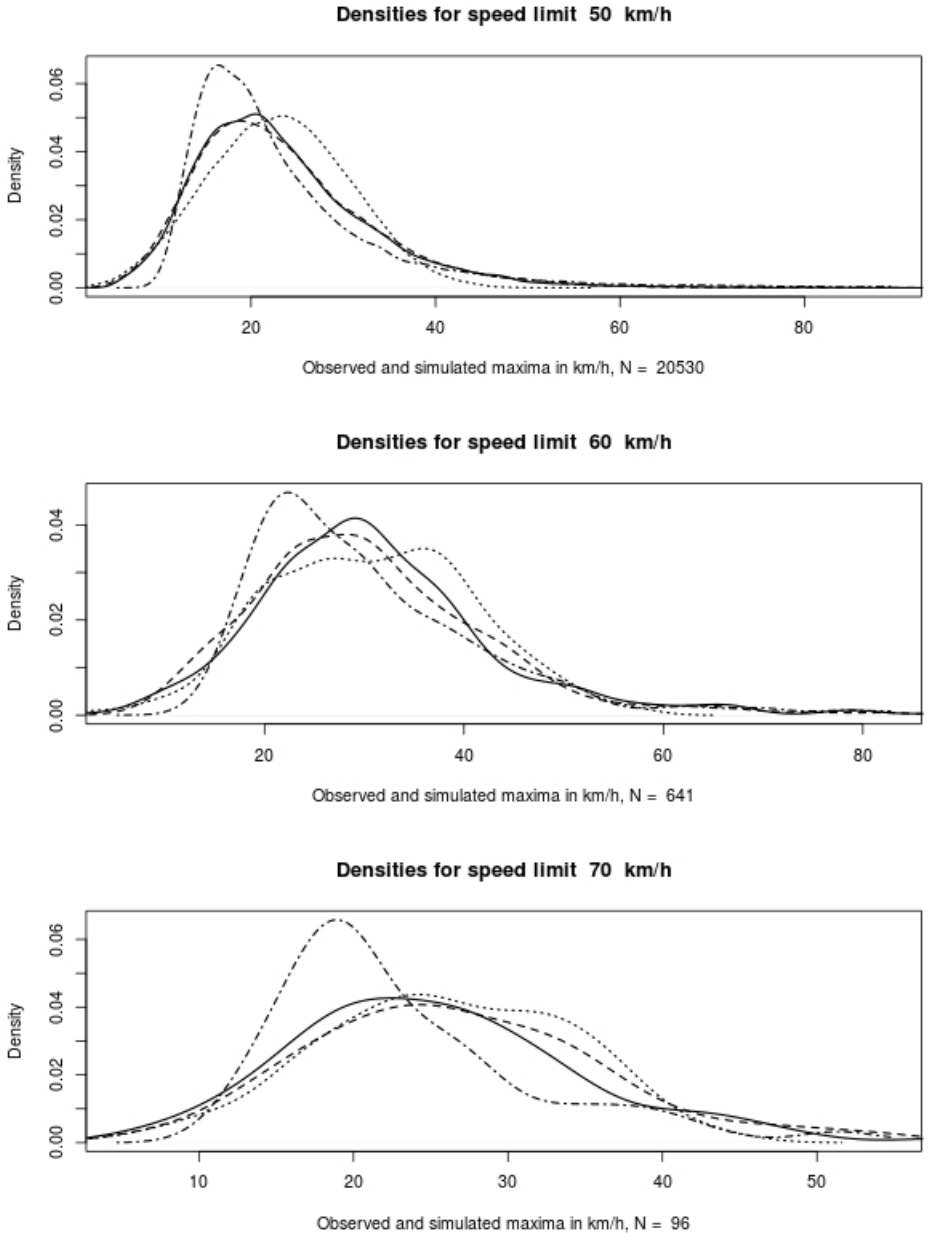

Figure A.5: Kernel density estimations for the speed limits 50, 60 and 70 km/h of the observed data (straight line), a simulated Weibull distribution (dotted line), a simulated Fréchet distribution (dotted-dashed line) and a simulated Gumbel distribution (dashed line). The parameters used for the simulation can be found in table 2.2 and 3.1. N denotes the number of observations used for the estimation. Observations on the x-axis are given in km/h.

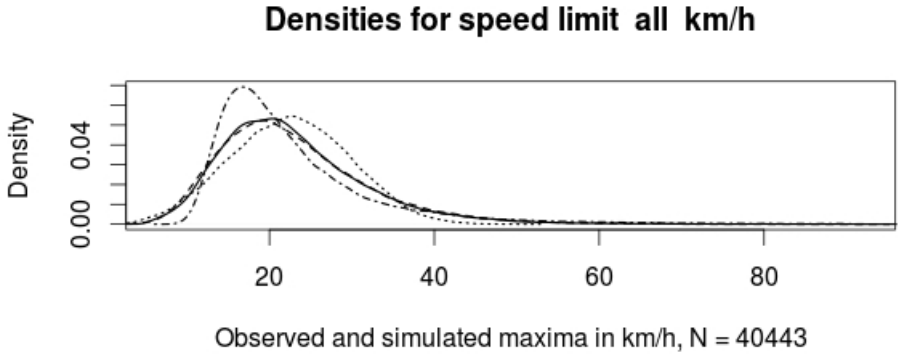

Figure A.6: Kernel density estimations for the pooled speed classes of the observed data (straight line), a simulated Weibull distribution (dotted line), a simulated Fréchet distribution (dotted-dashed line) and a simulated Gumbel distribution (dashed line). The parameters used for the simulation can be found in table 2.2 and 3.1. N denotes the number of observations used for the estimation. Observations on the x-axis are given in km/h.

A.3 Quantile-plots

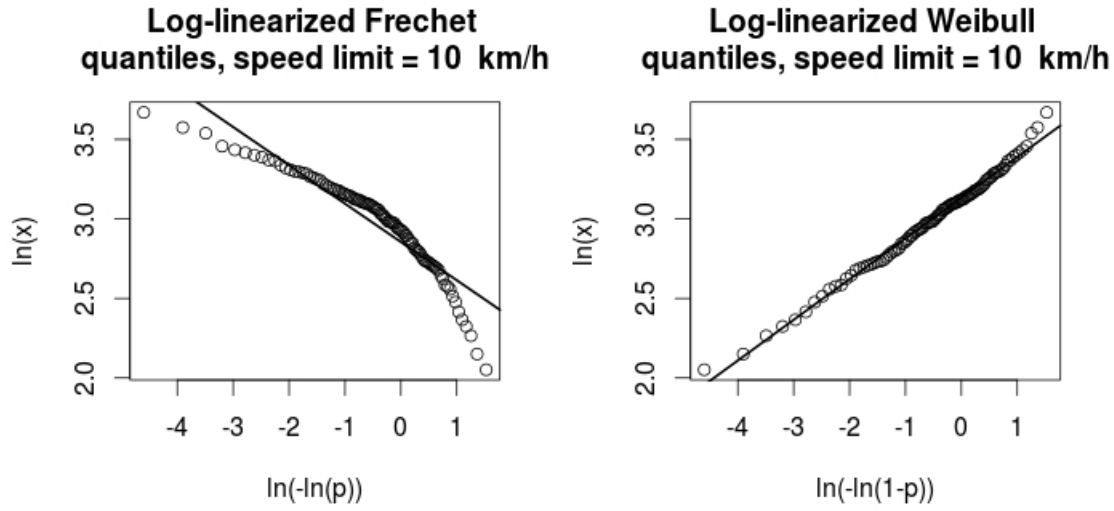

Figure A.7: Quantile-plots for the log-linearized quantiles by supposing a Fréchet or a Weibull distribution for the speed limit 10 km/h. The straight line represents the linear regression estimated from the data.

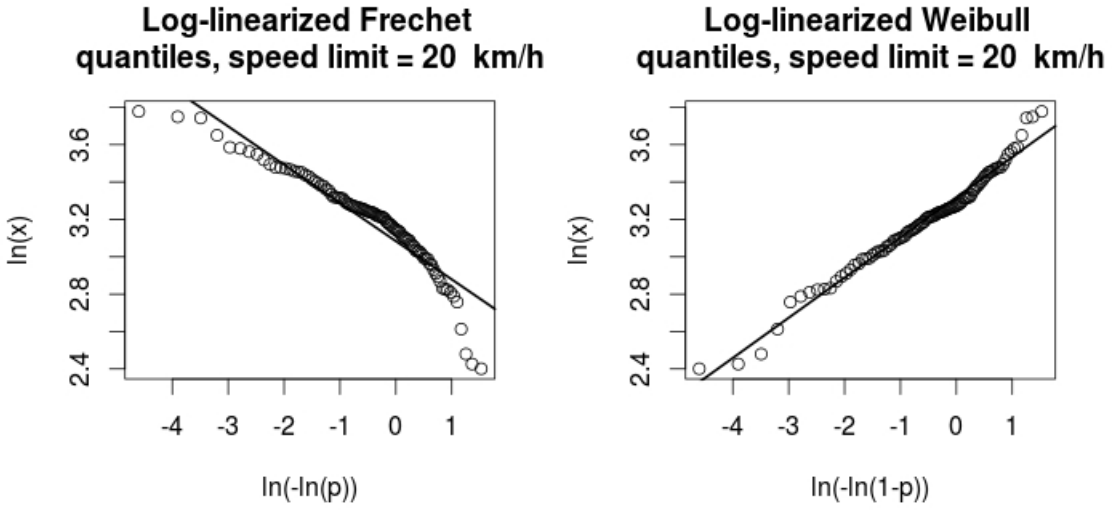

Figure A.8: Quantile-plots for the log-linearized quantiles by supposing a Fréchet or a Weibull distribution for the speed limit 20 km/h. The straight line represents the linear regression estimated from the data.

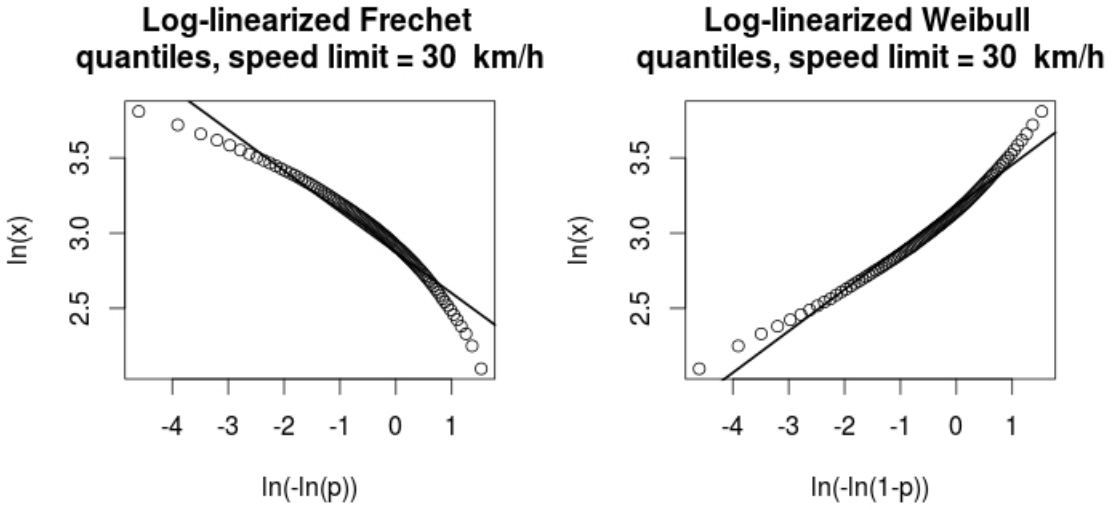

Figure A.9: Quantile-plots for the log-linearized quantiles by supposing a Fréchet or a Weibull distribution for the speed limit 30 km/h. The straight line represents the linear regression estimated from the data.

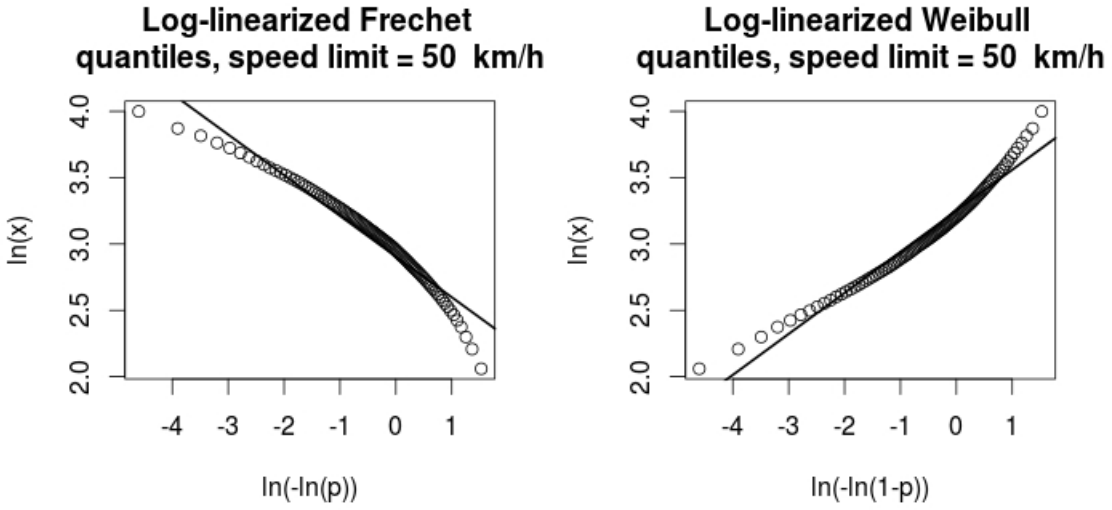

Figure A.10: Quantile-plots for the log-linearized quantiles by supposing a Fréchet or a Weibull distribution for the speed limit 50 km/h. The straight line represents the linear regression estimated from the data.

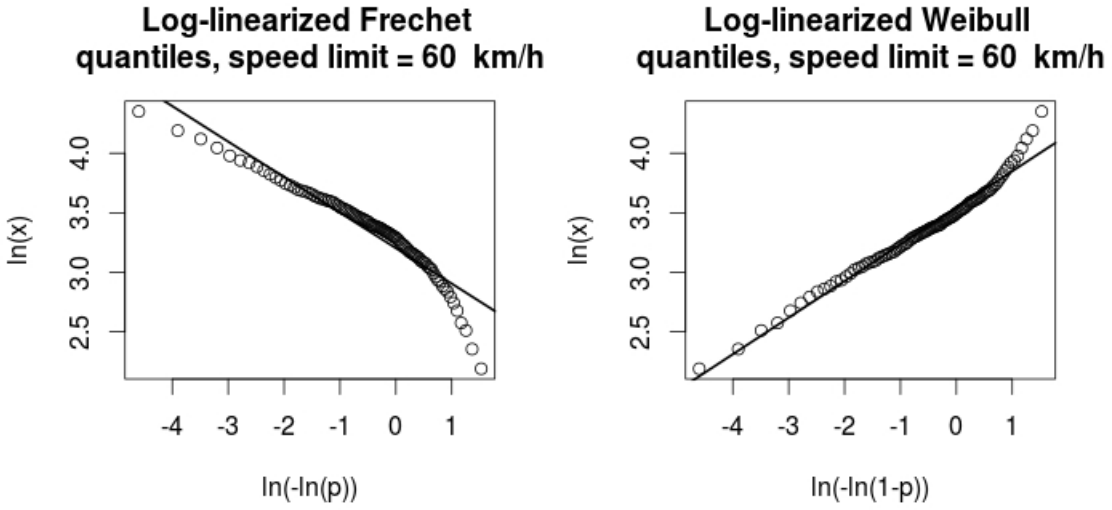

Figure A.11: Quantile-plots for the log-linearized quantiles by supposing a Fréchet or a Weibull distribution for the speed limit 60 km/h. The straight line represents the linear regression estimated from the data.

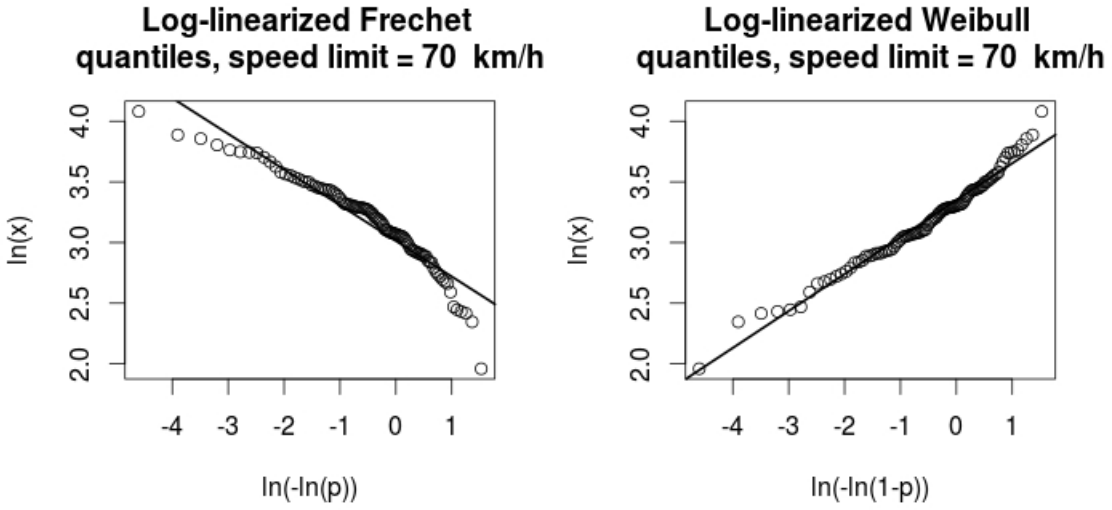

Figure A.12: Quantile-plots for the log-linearized quantiles by supposing a Fréchet or a Weibull distribution for the speed limit 70 km/h. The straight line represents the linear regression estimated from the data.

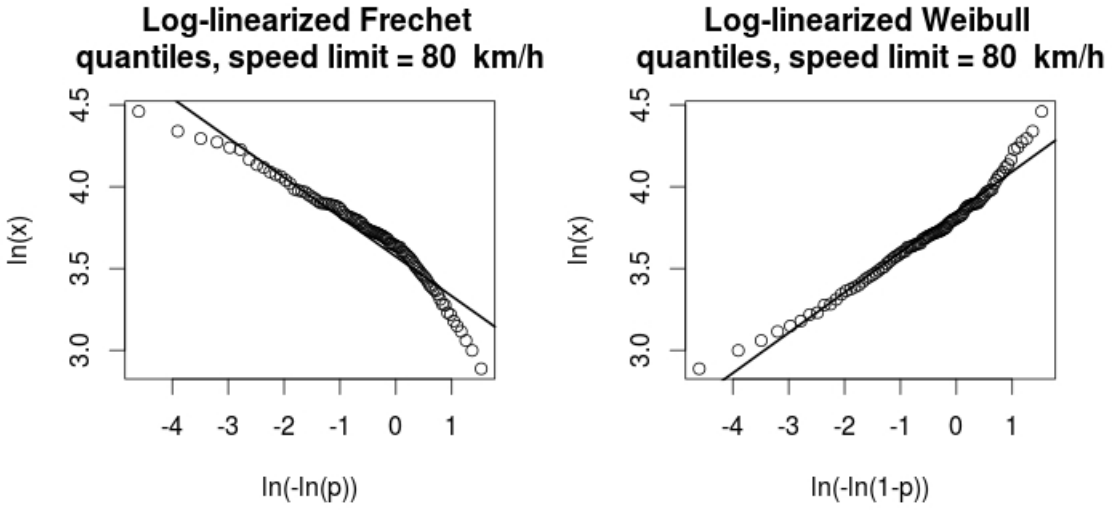

Figure A.13: Quantile-plots for the log-linearized quantiles by supposing a Fréchet or a Weibull distribution for the speed limit 70 km/h. The straight line represents the linear regression estimated from the data.

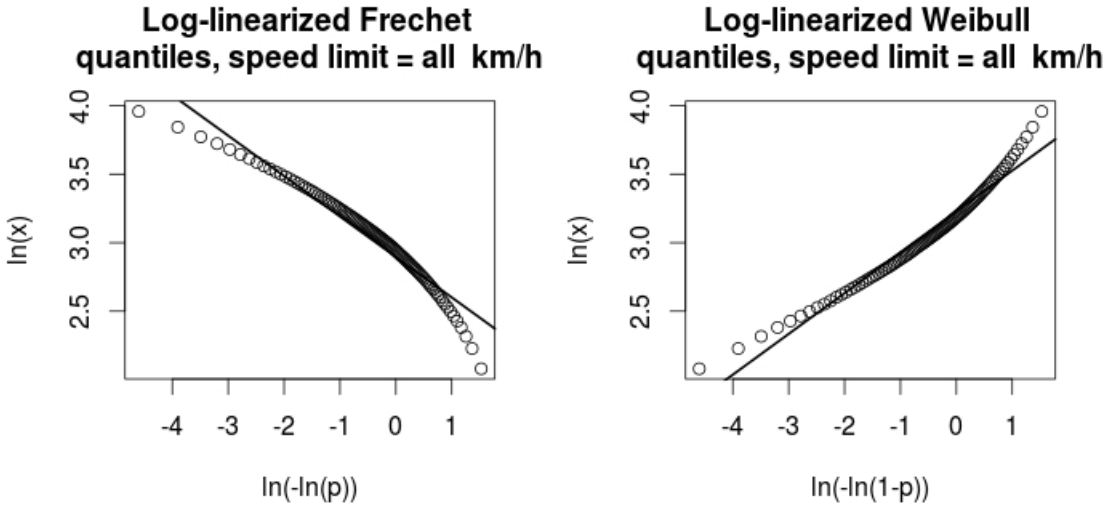

Figure A.14: Quantile-plots for the log-linearized quantiles by supposing a Fréchet or a Weibull distribution for all speed limits. The straight line represents the linear regression estimated from the data.

A.4 Quantile-quantile plots

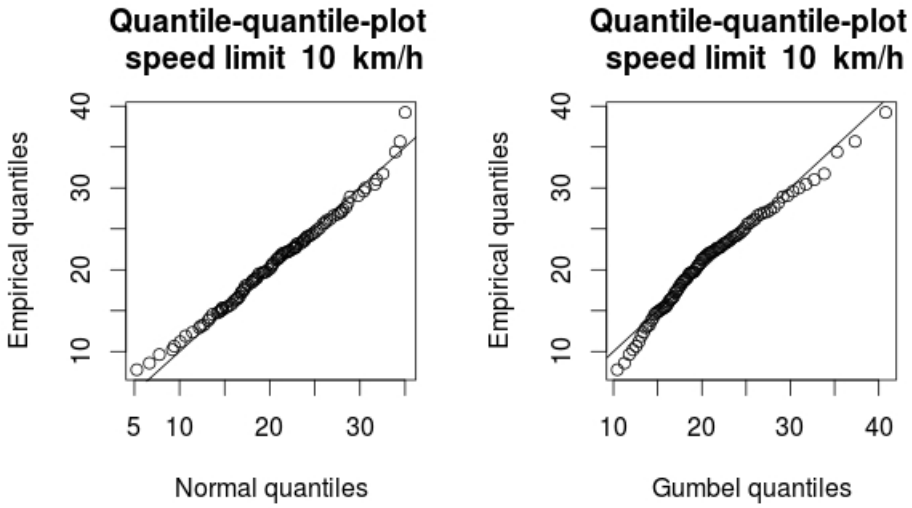

Figure A.15: Quantile-quantile-plots for the empirical quantiles and the theoretical quantiles of the Gaussian and the Gumbel distribution for the speed limit 10 km/h. The according parameters for the two distributions are estimated from the data and can be found in table 2.1 and 2.2.

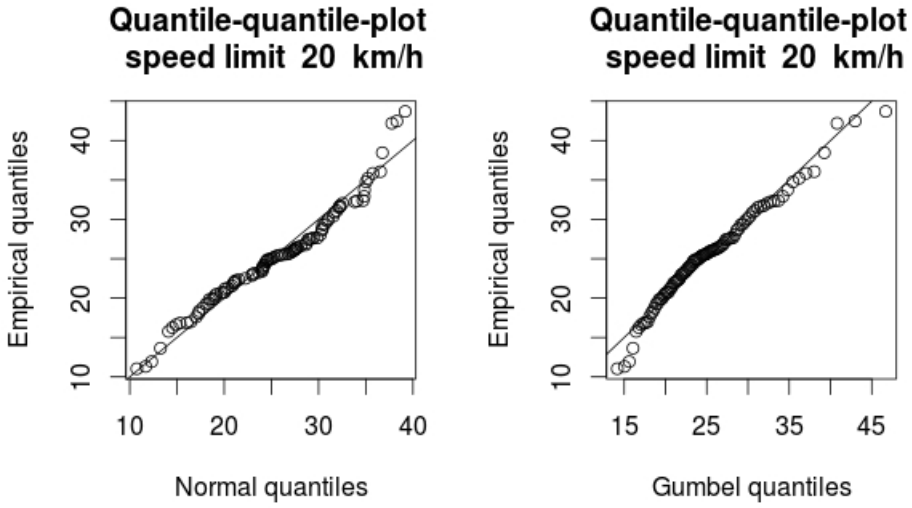

Figure A.16: Quantile-quantile-plots for the empirical quantiles and the theoretical quantiles of the Gaussian and the Gumbel distribution for the speed limit 20 km/h. The according parameters for the two distributions are estimated from the data and can be found in table 2.1 and 2.2.

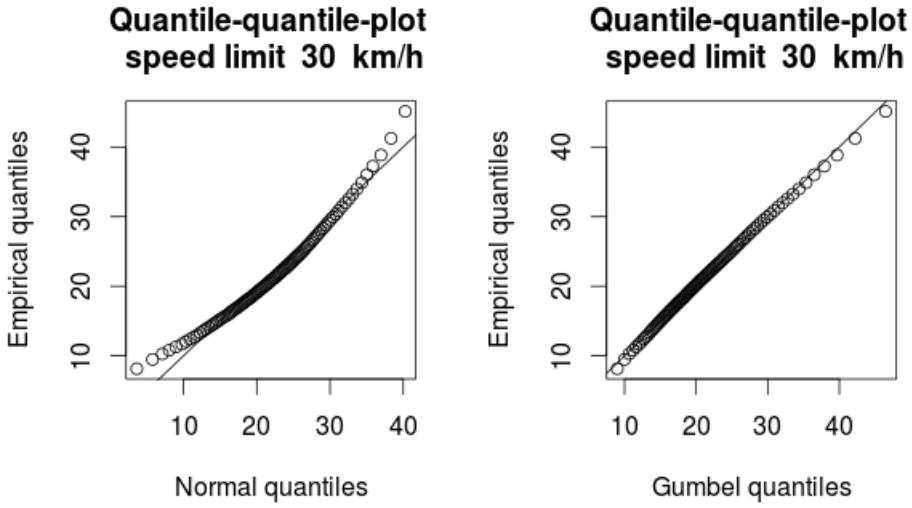

Figure A.17: Quantile-quantile-plots for the empirical quantiles and the theoretical quantiles of the Gaussian and the Gumbel distribution for the speed limit 30 km/h. The according parameters for the two distributions are estimated from the data and can be found in table 2.1 and 2.2.

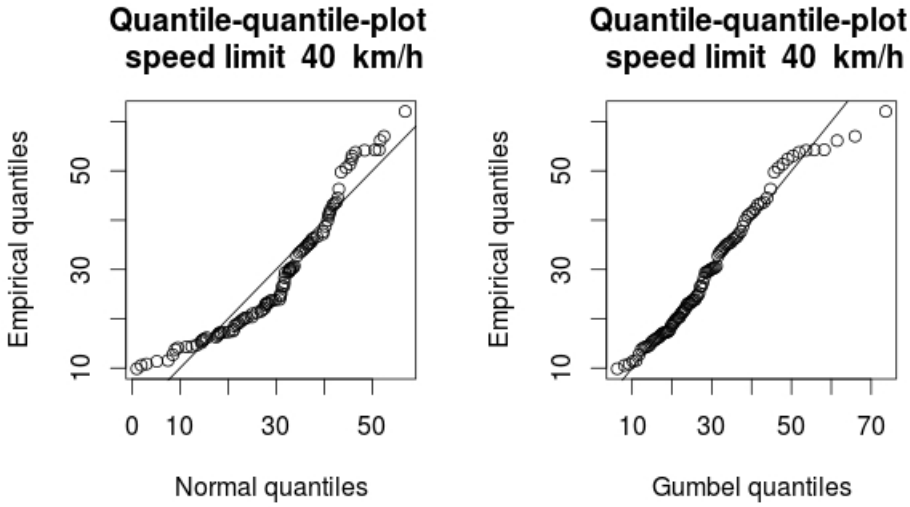

Figure A.18: Quantile-quantile-plots for the empirical quantiles and the theoretical quantiles of the Gaussian and the Gumbel distribution for the speed limit 40 km/h. The according parameters for the two distributions are estimated from the data and can be found in table 2.1 and 2.2.

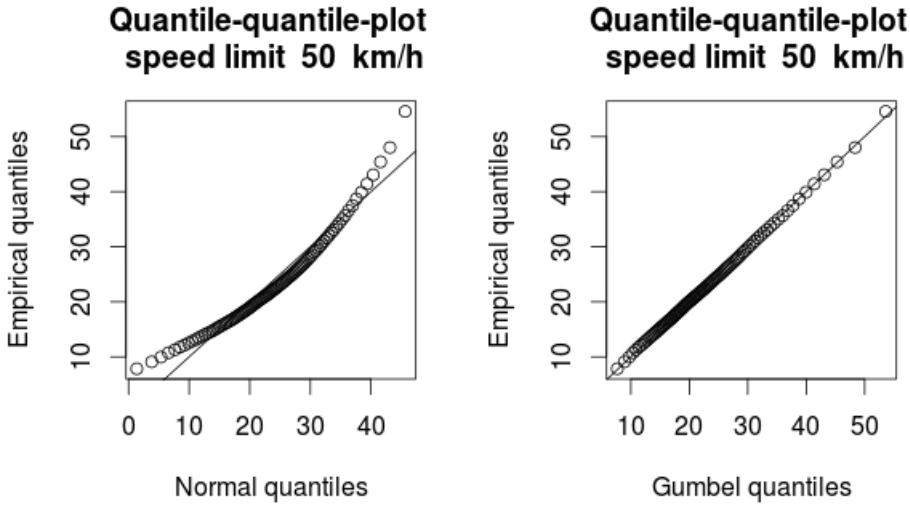

Figure A.19: Quantile-quantile-plots for the empirical quantiles and the theoretical quantiles of the Gaussian and the Gumbel distribution for the speed limit 50 km/h. The according parameters for the two distributions are estimated from the data and can be found in table 2.1 and 2.2.

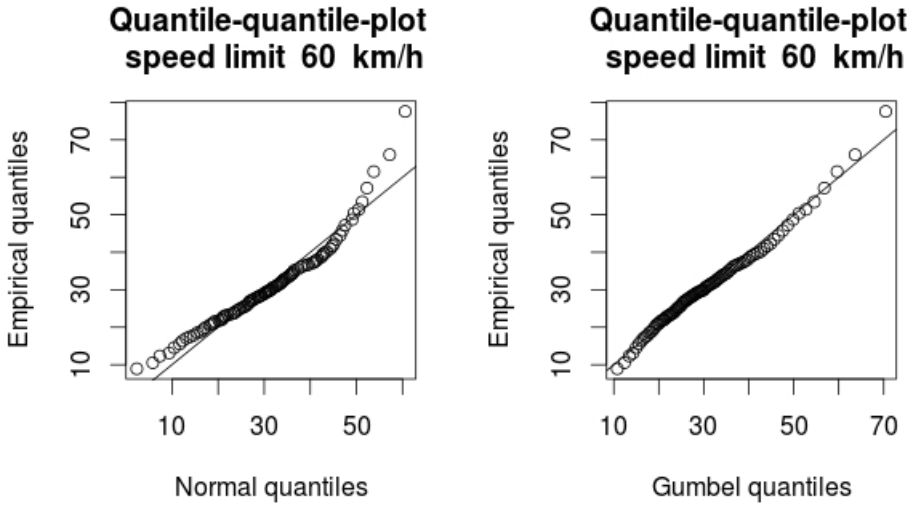

Figure A.20: Quantile-quantile-plots for the empirical quantiles and the theoretical quantiles of the Gaussian and the Gumbel distribution for the speed limit 60 km/h. The according parameters for the two distributions are estimated from the data and can be found in table 2.1 and 2.2.

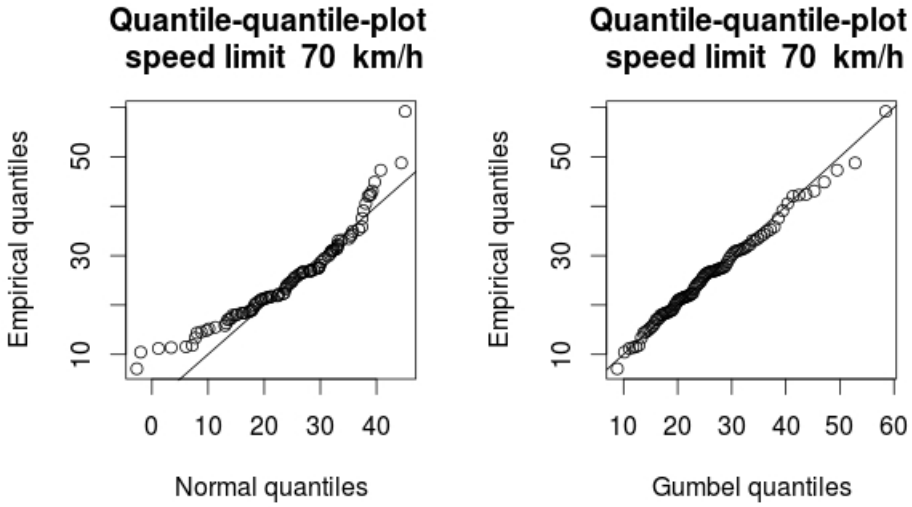

Figure A.21: Quantile-quantile-plots for the empirical quantiles and the theoretical quantiles of the Gaussian and the Gumbel distribution for the speed limit 70 km/h. The according parameters for the two distributions are estimated from the data and can be found in table 2.1 and 2.2.

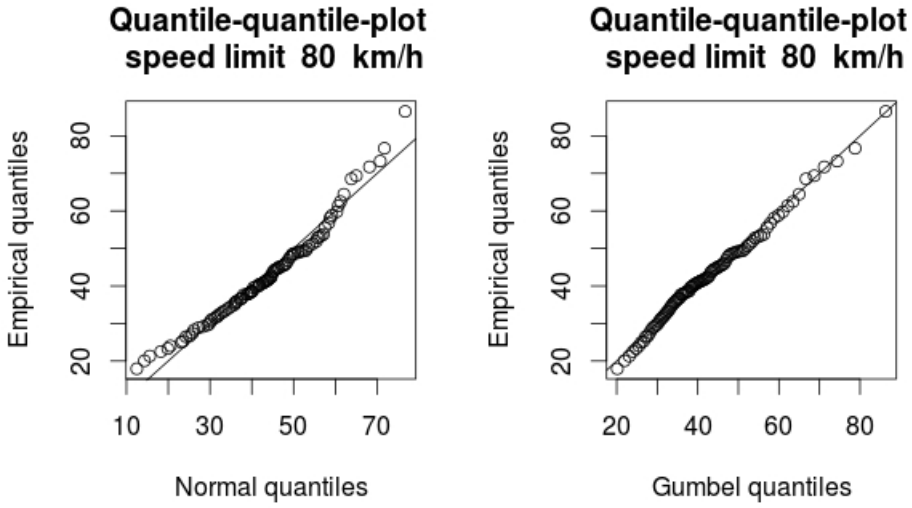

Figure A.22: Quantile-quantile-plots for the empirical quantiles and the theoretical quantiles of the Gaussian and the Gumbel distribution for the speed limit 80 km/h. The according parameters for the two distributions are estimated from the data and can be found in table 2.1 and 2.2.

A.5 Extreme value index estimations

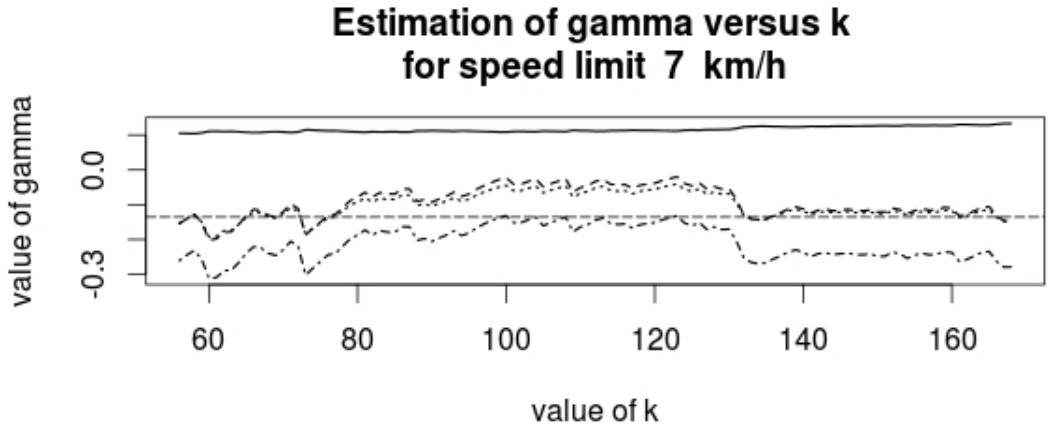

Figure A.23: Estimation of γ according to the range of the upper order statistics for the speed limit 7 km/h. The straight line represents Hill's estimator $\widehat{\gamma_1}$. The dashed line is the moment estimator $\widehat{\gamma_2}$. The dotted-dashed line represents the second moment estimator $\widehat{\gamma_3}$. And the dotted line represents the third moment estimator $\widehat{\gamma_4}$. The long dashed line is the estimator $\overline{\gamma}_{2,3,4}$. The according estimates can be found in table 5.1.

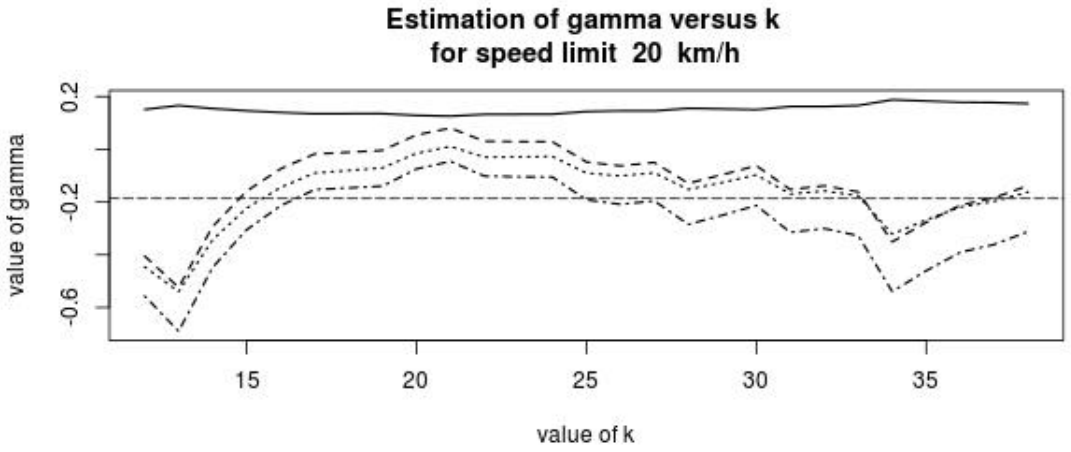

Figure A.24: Estimation of γ according to the range of the upper order statistics for the speed limit 20 km/h. The straight line represents Hill's estimator $\widehat{\gamma_1}$. The dashed line is the moment estimator $\widehat{\gamma_2}$. The dotted-dashed line represents the second moment estimator $\widehat{\gamma_3}$. And the dotted line represents the third moment estimator $\widehat{\gamma_4}$. The long dashed line is the estimator $\overline{\gamma}_{2,3,4}$. The according estimates can be found in table 5.1.

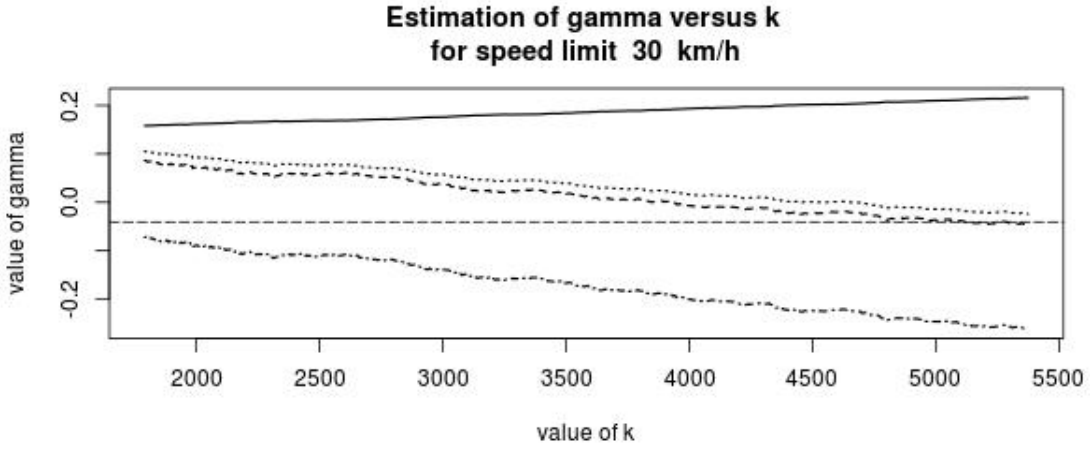

Figure A.25: Estimation of γ according to the range of the upper order statistics for the speed limit 30 km/h. The straight line represents Hill's estimator $\widehat{\gamma_1}$. The dashed line is the moment estimator $\widehat{\gamma_2}$. The dotted-dashed line represents the second moment estimator $\widehat{\gamma_3}$. And the dotted line represents the third moment estimator $\widehat{\gamma_4}$. The long dashed line is the estimator $\overline{\gamma}_{2,3,4}$. The according estimates can be found in table 5.1.

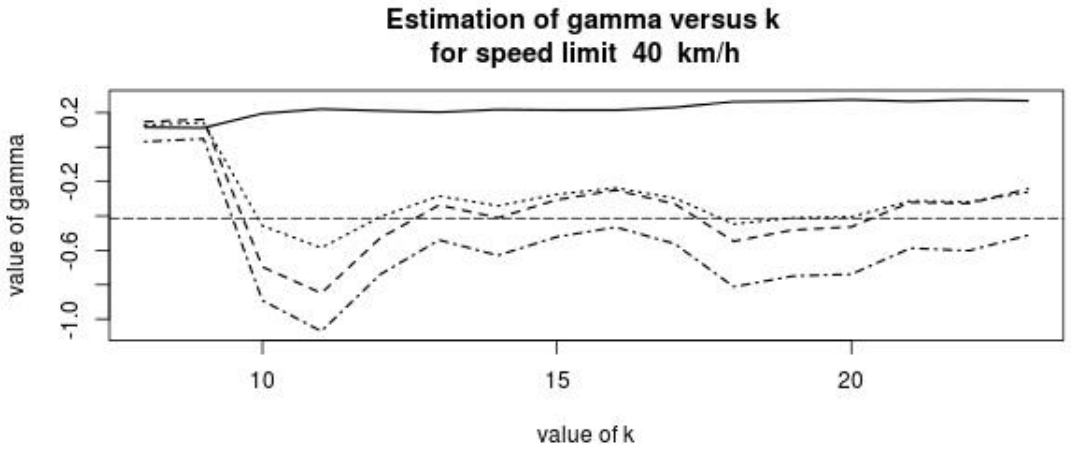

Figure A.26: Estimation of γ according to the range of the upper order statistics for the speed limit 40 km/h. The straight line represents Hill's estimator $\widehat{\gamma_1}$. The dashed line is the moment estimator $\widehat{\gamma_2}$. The dotted-dashed line represents the second moment estimator $\widehat{\gamma_3}$. And the dotted line represents the third moment estimator $\widehat{\gamma_4}$. The long dashed line is the estimator $\overline{\gamma}_{2,3,4}$. The according estimates can be found in table 5.1.

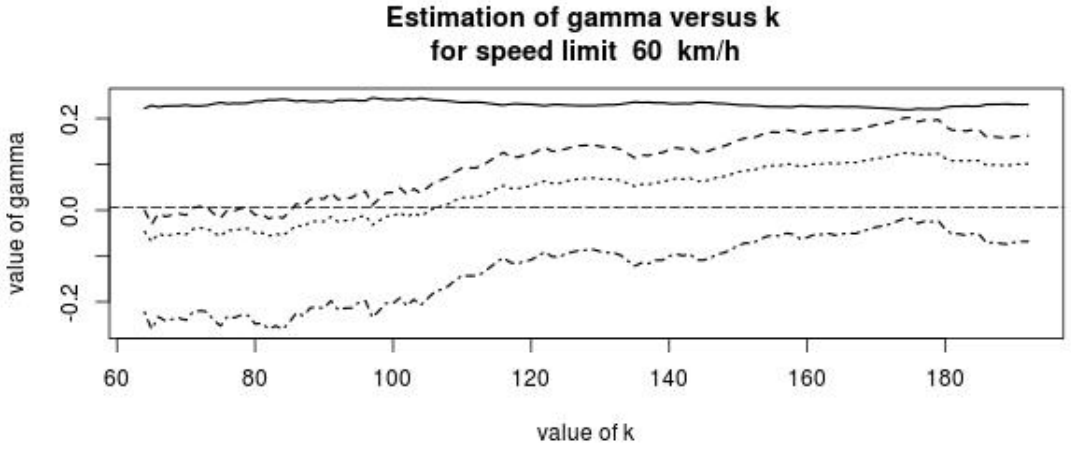

Figure A.27: Estimation of γ according to the range of the upper order statistics for the speed limit 60 km/h. The straight line represents Hill's estimator $\widehat{\gamma_1}$. The dashed line is the moment estimator $\widehat{\gamma_2}$. The dotted-dashed line represents the second moment estimator $\widehat{\gamma_3}$. And the dotted line represents the third moment estimator $\widehat{\gamma_4}$. The long dashed line is the estimator $\overline{\gamma}_{2,3,4}$. The according estimates can be found in table 5.1.

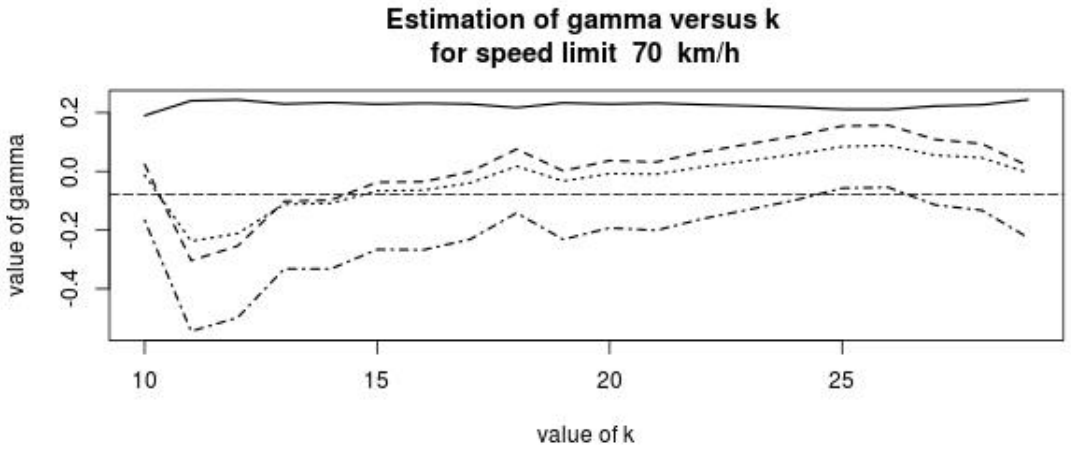

Figure A.28: Estimation of γ according to the range of the upper order statistics for the speed limit 70 km/h. The straight line represents Hill's estimator $\widehat{\gamma_1}$. The dashed line is the moment estimator $\widehat{\gamma_2}$. The dotted-dashed line represents the second moment estimator $\widehat{\gamma_3}$. And the dotted line represents the third moment estimator $\widehat{\gamma_4}$. The long dashed line is the estimator $\overline{\gamma}_{2,3,4}$. The according estimates can be found in table 5.1.

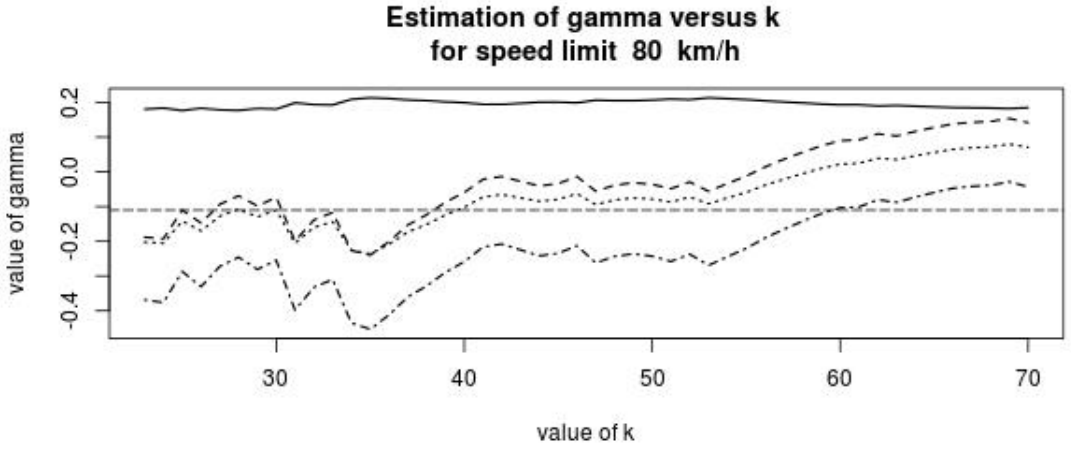

Figure A.29: Estimation of γ according to the range of the upper order statistics for the speed limit 80 km/h. The straight line represents Hill's estimator $\widehat{\gamma_1}$. The dashed line is the moment estimator $\widehat{\gamma_2}$. The dotted-dashed line represents the second moment estimator $\widehat{\gamma_3}$. And the dotted line represents the third moment estimator $\widehat{\gamma_4}$. The long dashed line is the estimator $\overline{\gamma}_{2,3,4}$. The according estimates can be found in table 5.1.

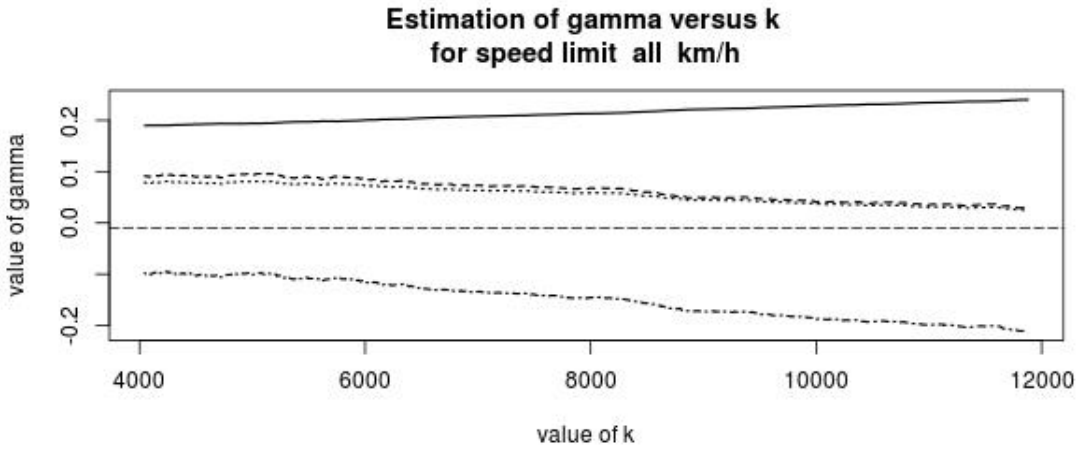

Figure A.30: Estimation of γ according to the range of the upper order statistics for all speed limits. The straight line represents Hill's estimator $\widehat{\gamma_1}$. The dashed line is the moment estimator $\widehat{\gamma_2}$. The dotted-dashed line represents the second moment estimator $\widehat{\gamma_3}$. And the dotted line represents the third moment estimator $\widehat{\gamma_4}$. The long dashed line is the estimator $\overline{\gamma}_{2,3,4}$. The according estimates can be found in table 5.1.

A.6 Endpoint estimations

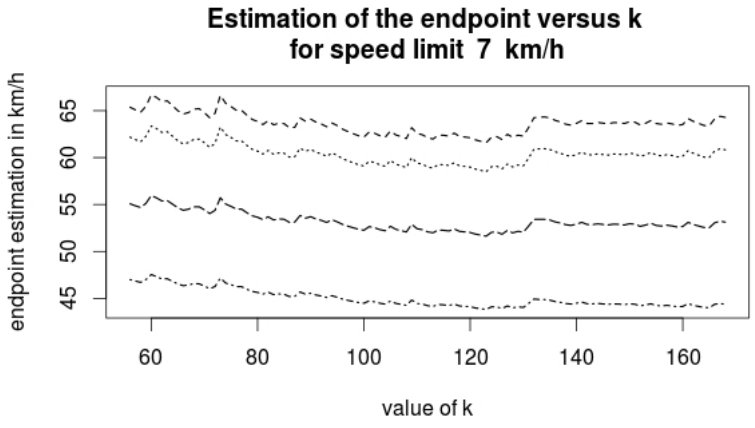

Figure A.31: Endpoint estimation in km/h of x^* according to the range of the upper order statistics for the speed class of 7 km/h. The dashed line represents the endpoint estimations using the moment estimator $\overline{\gamma_2}$. The dotted-dashed line represents endpoint estimations for the second moment estimator $\overline{\gamma_3}$. And the dotted line represents the endpoints for the third moment estimator $\overline{\gamma_4}$. The long dashed line is the endpoint obtained by the estimator $\overline{\gamma}_{2,3,4}$. The according estimates can be found in table 5.4.

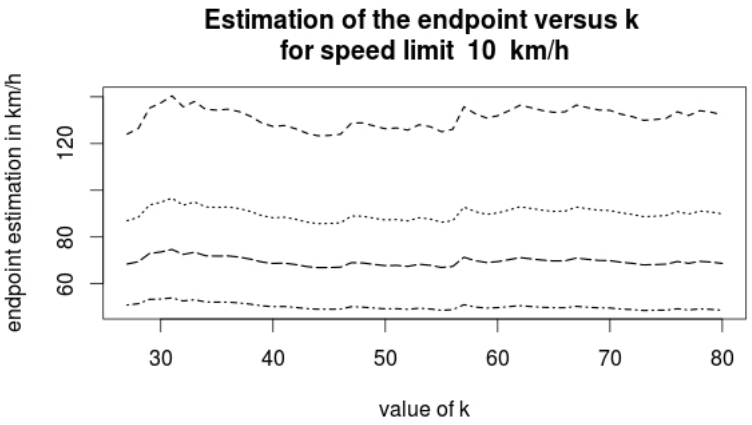

Figure A.32: Endpoint estimation in km/h of x^* according to the range of the upper order statistics for the speed class of 10 km/h. The dashed line represents the endpoint estimations using the moment estimator $\overline{\gamma_2}$. The dotted-dashed line represents endpoint estimations for the second moment estimator $\overline{\gamma_3}$. And the dotted line represents the endpoints for the third moment estimator $\overline{\gamma_4}$. The long dashed line is the endpoint obtained by the estimator $\overline{\gamma}_{2,3,4}$. The according estimates can be found in table 5.4.

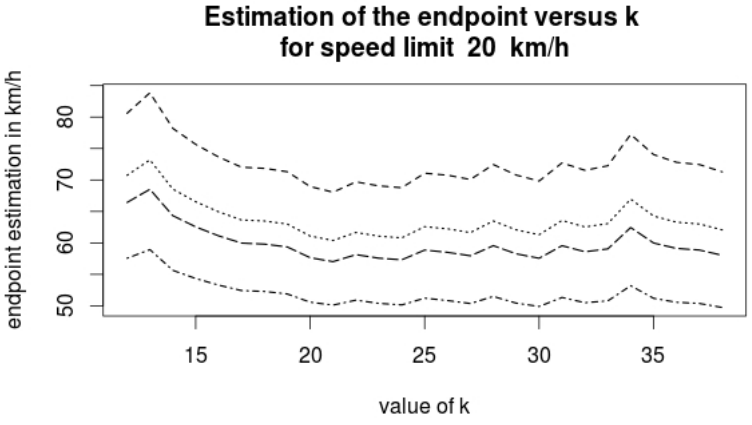

Figure A.33: Endpoint estimation in km/h of x^* according to the range of the upper order statistics for the speed class of 20 km/h. The dashed line represents the endpoint estimations using the moment estimator $\overline{\gamma_2}$. The dotted-dashed line represents endpoint estimations for the second moment estimator $\overline{\gamma_3}$. And the dotted line represents the endpoints for the third moment estimator $\overline{\gamma_4}$. The long dashed line is the endpoint obtained by the estimator $\overline{\gamma}_{2,3,4}$. The according estimates can be found in table 5.4.

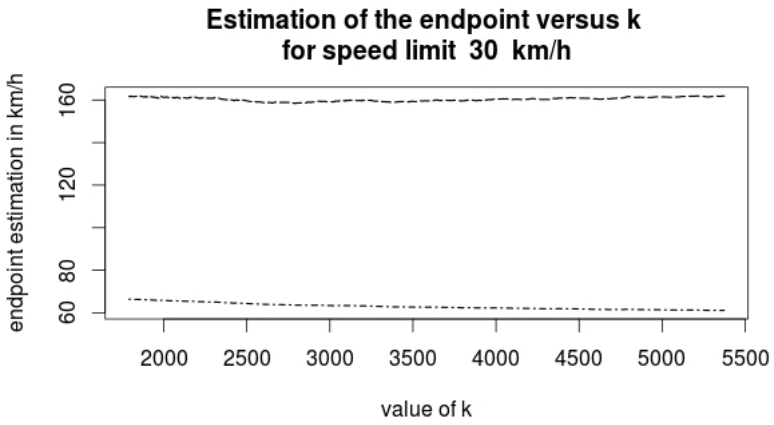

Figure A.34: Endpoint estimation in km/h of x^* according to the range of the upper order statistics for the speed class of 30 km/h. The dotted-dashed line represents endpoint estimations for the second moment estimator $\overline{\gamma_3}$. The long dashed line is the endpoint obtained by the estimator $\overline{\gamma}_{2,3,4}$. The according estimates can be found in table 5.4.

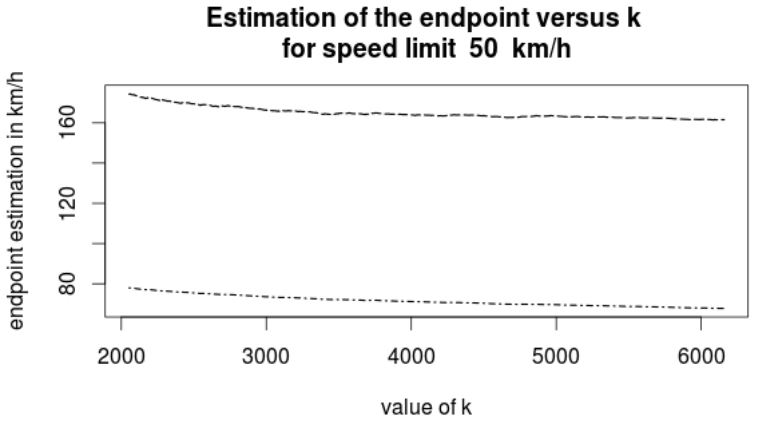

Figure A.35: Endpoint estimation in km/h of x^* according to the range of the upper order statistics for the speed class of 50 km/h. The dotted-dashed line represents endpoint estimations for the second moment estimator $\overline{\gamma_3}$. The long dashed line is the endpoint obtained by the estimator $\overline{\gamma}_{2,3,4}$. The according estimates can be found in table 5.4.

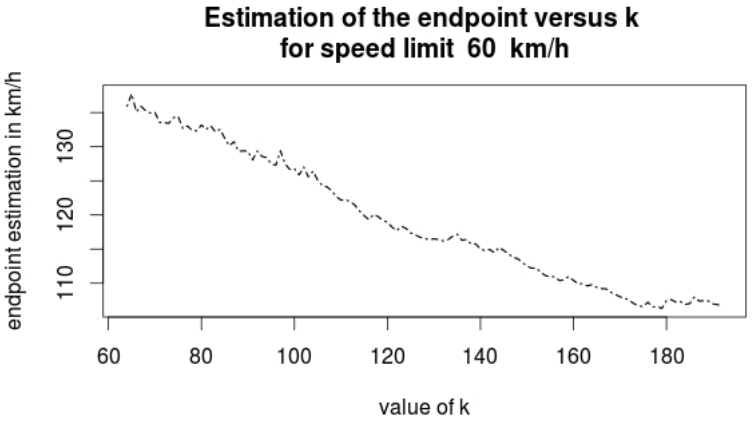

Figure A.36: Endpoint estimation in km/h of x^* according to the range of the upper order statistics for the speed class of 60 km/h. The dotted-dashed line represents endpoint estimations for the second moment estimator $\overline{\gamma_3}$. The according estimates can be found in table 5.4.

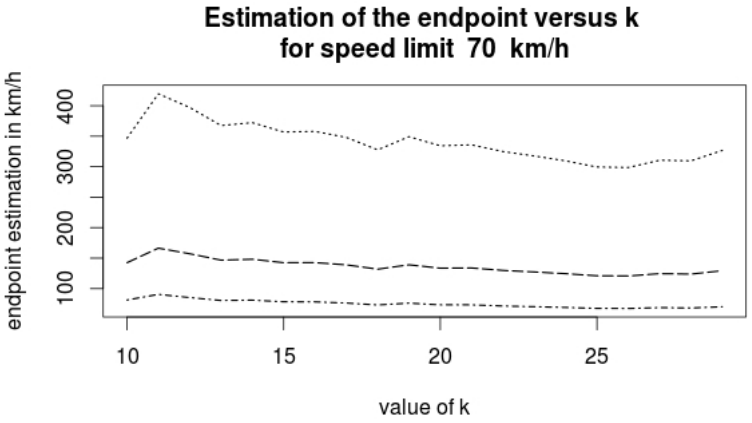

Figure A.37: Endpoint estimation in km/h of x^* according to the range of the upper order statistics for the speed class of 70 km/h. The dotted-dashed line represents endpoint estimations for the second moment estimator $\overline{\gamma_3}$. And the dotted line represents the endpoints for the third moment estimator $\overline{\gamma_4}$. The long dashed line is the endpoint obtained by the estimator $\overline{\gamma}_{2,3,4}$. The according estimates can be found in table 5.4.

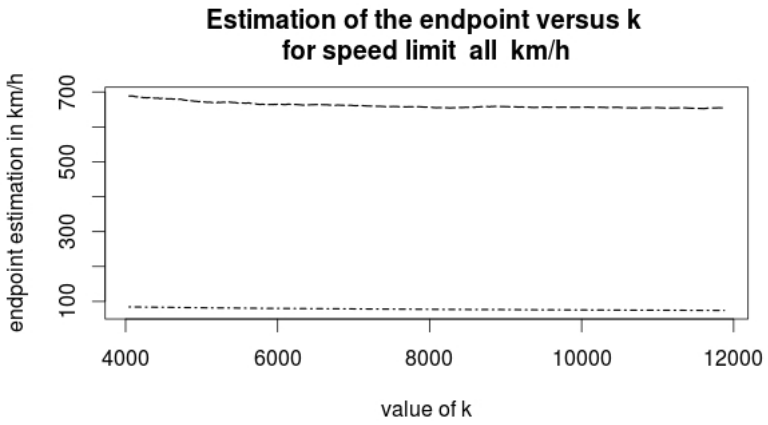

Figure A.38: Endpoint estimation in km/h of x^* according to the range of the upper order statistics for all speed limits. The dotted-dashed line represents endpoint estimations for the second moment estimator $\overline{\gamma_3}$. The long dashed line is the endpoint obtained by the estimator $\overline{\gamma}_{2,3,4}$. The according estimates can be found in table 5.4.

Declaration of Authorship

I hereby confirm that I have authored this Bachelor's thesis independently and without use of others than the indicated sources. All passages which are literally or in general matter taken out of publications or other sources are marked as such.

Paris, 4th October, 2013

Alexander Buchholz